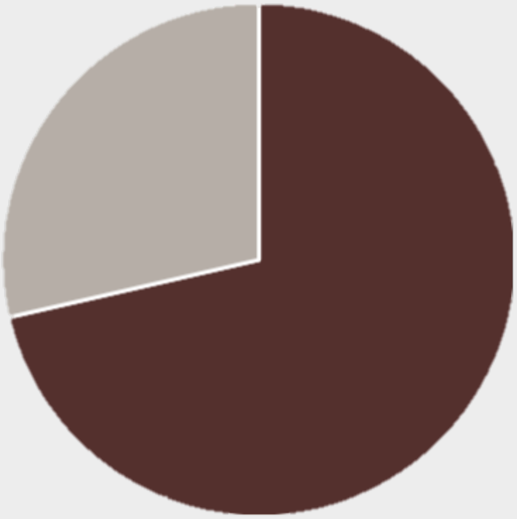
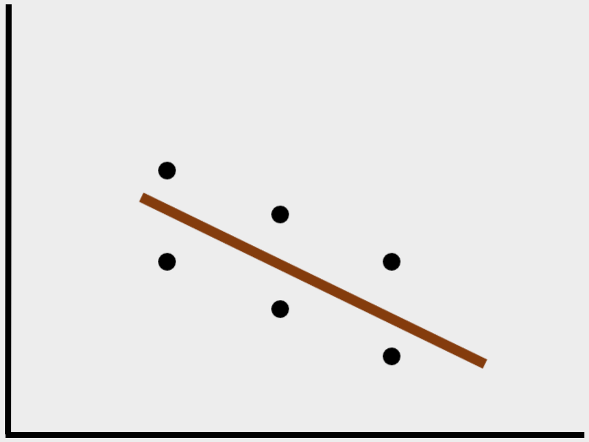
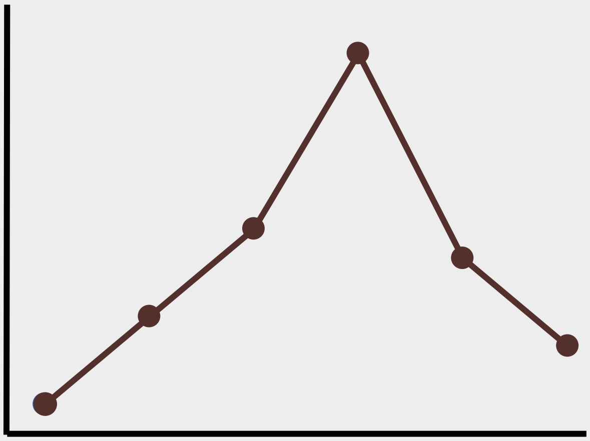
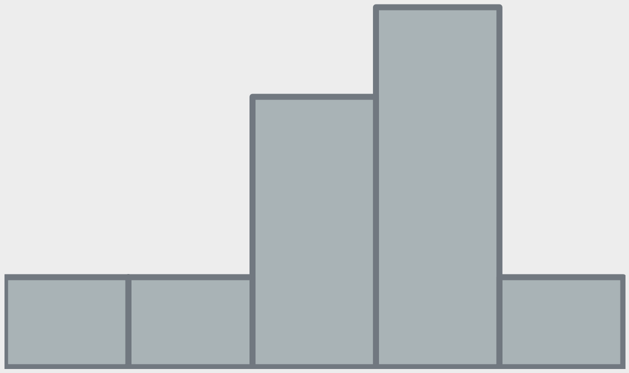
SIGNIFICANT STATISTICS

AN INTRODUCTION TO STATISTICS









ADAPTED BY JOHN MORGAN RUSSELL

Statistics is about separating the signal from the noise, deciphering what is actually significant versus what is just happening due to random chance. In addition to demonstrating the basic concepts needed to do that, ***Significant Statistics: An Introduction to Statistics*** attempts to focus on what is significant and eliminate some of the noise that may commonly be found in many introductory statistics texts.

This book is intended for a one-semester introduction to statistics course for students who are not mathematics or engineering majors. It focuses on the interpretation of statistical results, especially in real world settings, and assumes that students have an understanding of intermediate algebra. It covers the basic ideas of data collection, univariate and bivariate descriptives, and one and two sample inference for means and proportions. It does not emphasize one specific software, and instead offers links to many extra resources. Examples of each topic are explained step-by-step throughout the text and followed by 'Your Turn' problems that are designed as extra practice for students. Additionally, there are over 600 practice problems at the end of the book.

COLLEGE OF SCIENCE
STATISTICS
VIRGINIA TECH

In association with

VIRGINIA TECH
PUBLISHING

DOI: https://doi.org/10.21061/significantstatistics

ISBN (PDF): 978-1-962841-01-6

Significant Statistics: An Introduction to Statistics

Significant Statistics

An Introduction to Statistics

Adapted by

John Morgan Russell

From
Barbara Illowsky and Susan Dean
and
David Diez, Mine Cetinkaya–Rundel, and Christopher D. Barr
and
Julie Vu and David Harrington

PDF AND EPUB FREE ONLINE AT:
https://doi.org/10.21061/significantstatistics

COLLEGE OF SCIENCE
STATISTICS
VIRGINIA TECH

VIRGINIA TECH
PUBLISHING

This work is published by the Virginia Tech Department of Statistics in association with Virginia Tech Publishing and the Open Education Initiative.
BLACKSBURG, VA

Attributions: The base of the book is from *OpenStax Introductory Statistics* by Barbara Illowsky and Susan Dean, which is licensed with a Creative Commons Attribution 4.0 (CC BY 4.0) license, much of which was reworded and reorganized. Additional content from *OpenIntro Statistics* by David Diez, Mine Cetinkaya-Rundel, and Christopher D. Barr and *Introductory Statistics for the Life and Biomedical Sciences* by Julie Vu and David Harrington, both licensed with a Creative Commons Attribution Share-Alike 3.0 (CC BY-SA 3.0) license, were then added to fill in gaps. Several figures were also adapted from the OpenIntro texts.

Suggested citation: Russell, John Morgan (2025). *Significant Statistics: An Introduction to Statistics*. Blacksburg: Virginia Tech Department of Statistics. https://doi.org/10.21061/significantstatistics. Licensed with CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0.

Publisher: This work is published by the Virginia Tech Department of Statistics in association with Virginia Tech Publishing and the Open Education Initiative, a division of the University Libraries at Virginia Tech.

Virginia Tech Department of Statistics
Cheatham Hall
310 West Campus Drive
Blacksburg, VA 24061

Virginia Tech Publishing
University Libraries at Virginia Tech
560 Drillfield Drive
Blacksburg, VA 24061
publishing@vt.edu

Open Education Initiative
University Libraries at Virginia Tech
560 Drillfield Drive
Blacksburg, VA 24061
openeducation@vt.edu

Peer review: This book has undergone single-anonymous peer review by two reviewers.

Accessibility statement: Virginia Tech is committed to making its publications accessible in accordance with the Americans with Disabilities Act of 1990. The text, images, and links in the PDF version of this text are tagged structurally and include alternative text, which allows for machine readability. We are continuously working to improve accessibility and welcome any feedback from readers.

Publication cataloging information:
Russell, John Morgan, author
Significant Statistics: An Introduction to Statistics / John Morgan Russell
Pages cm
ISBN 978-1-962841-01-6 (PDF)
ISBN 978-1-962841-00-9 (EPUB)
ISBN 978-1-957213-85-9 (Pressbooks; https://pressbooks.lib.vt.edu/significantstatistics)
ISBN 978-1-957213-86-6 (Print)
URI: https://hdl.handle.net/10919/118481
DOI: https://doi.org/10.21061/significantstatistics

Statistics--Textbooks
QA276.12 .I45 2025

Cover design: Kindred Grey

Contents

Chapter 1: Sampling and Data

Chapter 2: Univariate Descriptive Statistics

Chapter 3: Bivariate Descriptive Statistics

Chapter 4: Probability Distributions

Chapter 5: Foundations of Inference

Chapter 6: Inference for One Sample

Chapter 7: Inference for Two Samples

Extra Practice

Acknowledgments

Editorial Team

Anita Walz is Associate Professor and Assistant Director of Open Education and Scholarly Communication Librarian in the University Libraries at Virginia Tech. She received her MS in Library and Information Science from the University of Illinois at Urbana-Champaign and has worked in university, government, school, and international libraries for eighteen years. She is the founder of the Open Education Initiative at Virginia Tech and the managing editor of over twenty open textbooks adapted or created at Virginia Tech, many of which may be found by visiting VTechWork's Open Textbook Collection. For *Significant Statistics*, she provided overall planning, project coordination, day-to-day supervision, and oversight.

Kindred Grey is the Graphic Design and Publishing Specialist in the University Libraries at Virginia Tech. She joined University Libraries after receiving her BS in Statistics and Psychology from Virginia Tech in 2020. Kindred has contributed to over twenty open textbooks, providing technical support on layout and design. Her main focus is publishing open textbooks that are visually appealing, accessible, student oriented, and technologically advanced.

Reviewers

This resource has undergone single-anonymous peer review. Thank you to **Ilhan Izmirli, PhD** (*Associate Professor, Department of Statistics, George Mason University*) and **Katherine Bowe** (*Assistant Professor of Mathematics, Concord University*) for taking time to review and comment on this resource, ensuring its usefulness as a learning tool for instructors and students.

Project Funding

Development of this book was made possible in part through a grant from the University Libraries at Virginia Tech's Open Education Initiative.

Sources

Significant Statistics: An Introduction to Statistics is adapted by John Morgan Russell from open textbooks from OpenStax and OpenIntro. These source books are released under open licenses that allow reuse and remix at no cost with attribution. Additional topics, examples, and innovations in terminology and practical applications have been added, all with a goal of increasing relevance and accessibility for students.

Content for this book was gathered and adapted from multiple openly-licensed sources:

- OpenStax *Introductory Statistics* by Barbara Illowsky and Susan Dean, which is licensed with a Creative Commons Attribution 4.0 (CC BY 4.0) license
- *OpenIntro Statistics* by David Diez, Mine Cetinkaya-Rundel, and Christopher D. Barr, which is licensed with a Creative Commons Attribution Share-Alike 3.0 (CC BY SA 3.0) license
- *Introductory Statistics for the Life and Biomedical Sciences* by Julie Vu and David Harrington, which is licensed with a Creative Commons Attribution Share-Alike 3.0 (CC BY SA 3.0) license

The base of the book is from OpenStax, much of which was reworded and reorganized. The main reorganizations involved streamlining the probability chapter (Chapter 3), removing ancillary discrete (Chapter 4) and continuous (Chapter 5) distribution sections, introducing normal distribution (5.3) in the continuous distributions chapter, and removing the chi-square and ANOVA chapters.

Additional content from the OpenIntro texts was then added to fill in gaps. This included adding more detailed information on data collection (1.3 & 1.4), normal approximation (5.3), and inferential techniques applied to proportions (7.3). Several figures were also adapted from the OpenIntro texts.

Introduction

Welcome to *Significant Statistics: An Introduction to Statistics*. This textbook was written to provide students access to high-quality learning materials at no cost. These types of materials available under Creative Commons licenses are often called open educational resources (OER).

Statistics is about separating the signal from the noise, differentiating between actual significance and random chance occurrences. In addition to demonstrating the basic concepts underlying that task, this book attempts to focus on what is significant, eliminating some of the noise commonly found in introductory statistics texts.

In this book, I have "remixed" sections from two of the most widely used OER texts in the introductory statistics space and sprinkled in some thoughts of my own. This book does not lean on any specific technology, focusing instead on concepts.

Most sections feature worked examples with solutions, some of which have interactive features. Then "Your Turn!" problems are included to encourage readers to try similar problems independently. Each end-of-chapter "Wrap-Up" includes a chapter quiz, list of key terms, and links to practice problems and other resources.

Thanks for choosing this book. I hope it proves useful!

Sincerely,

John Morgan Russell

Significant Statistics: An Introduction to Statistics is intended as a one-semester introduction to statistics for students who are not mathematics or engineering majors. It focuses on the interpretation of statistical results, especially in real-world settings, and assumes that students have an understanding of intermediate algebra. In addition to plenty of practice problems, examples of each topic are explained step by step throughout the text.

Having successfully completed the course, the student should be able to:

- Identify and critique the use of statistical reasoning in science, industry, and public discourse
- Identify the appropriate data needed to answer research questions
- Assign appropriate data collection methods
- Apply appropriate methods of data visualization to explore data from a variety of disciplines
- Analyze provided data and use relevant technology when needed
- Appropriately interpret results of data exploration and statistical tests
- Employ critical thinking to make decisions
- Apply ethical reasoning and principles to scientific research

This book will cover the following topics:

- Chapter 1: Sampling and Data
- Chapter 2: Univariate Descriptive Statistics
- Chapter 3: Bivariate Descriptive Statistics
- Chapter 4: Probability Distributions
- Chapter 5: Foundations of Inference
- Chapter 6: Inference for One Sample
- Chapter 7: Inference for Two Samples

About the Author

John Morgan Russell teaches various introductory statistics courses at Virginia Tech and previously taught at George Mason University and Old Dominion University. He earned a BS in Mathematics from Christopher Newport University, an MS in Statistical Science from George Mason University, and an EdS in Instructional Design and Technology from Virginia Tech. His interests include statistics education, instructional design, and open educational resources.

Instructor Resources

Dear Colleague,

We all know the issues with the price of higher education. One small but still significant aspect over which instructors may have some level of control is the materials used. The use of open educational resources (OER) is a growing trend that I hope will continue to catch on.

In the realm of introductory statistics, there are many OER options available, the most complete being *Introductory Statistics* from OpenStax and *OpenIntro Statistics*. While these are both adequate options, the beauty of OER is that you can customize material to the needs of your course and students, which is what I have tried to do here. Specific to this book, I have "remixed" sections from the aforementioned texts and sprinkled in some thoughts of my own.

The intended audience for *Significant Statistics: An Introduction to Statistics* includes students who may not have completed a calculus prerequisite. In contrast to similar introductory statistics texts, this text has an emphasis on data collection as well as univariate and bivariate descriptives but does not deeply delve into probability topics. The ideas of simple linear regression are treated non-inferentially and covered early in the chapter dedicated to bivariate data (Chapter 3). The text takes a repetitive approach to one-sample inference to drive those basic concepts home. It ends with introductory level material on two-sample inference (Chapter 7). Please see the introduction for more details.

I've also tried to make this book technology-agnostic in order to accommodate a wide variety of preferences. It is my belief that if students understand the concepts, they can figure out how to use any technology to accomplish the task at hand. However, if learners rely on technology when first introduced to the concepts, true understanding may not be achieved.

I hope you will consider using this text!

Sincerely,

John Morgan Russell

About Open Educational Resources (OER)

OER are free teaching and learning materials that are licensed to allow for revision and reuse. They can be fully self-contained textbooks, videos, quizzes, learning modules, and more. OER are distinct from free online resources in that they permit others to use, copy, distribute, modify, or reuse the content. The legal permission to modify and customize OER to meet courses' specific learning objectives make them useful pedagogical tools.

About Pressbooks

Pressbooks is an open source, web-based authoring tool based on WordPress. Virginia Tech personnel use it to create and adapt openly licensed course materials. Pressbooks offers an intuitive editing environment for educators interested in customizing this book.

About This Book

Significant Statistics: An Introduction to Statistics is an open textbook. You are welcome to freely use, adapt, and share this book with attribution according to the Creative Commons Attribution ShareAlike 4.0 (CC BY-SA 4.0) license. This license allows customization and redistribution for any purpose, even commercially as long as attribution is made, and derivative contributions are distributed under the same license. See links for the license deed and recommended practices for attribution below. All figures in this book are licensed under Creative Commons licenses or used under fair use. Please see the end-of-chapter references for license information for each figure before reuse.

Supplemental multimedia material aligned with this textbook, including videos, audio-only versions of the videos in podcast format, and PowerPoint lecture notes, can be found at:

- Significant Statistics Website
- Significant Statistics YouTube Channel
- Significant Statistics Podcast Channel

Please use the instructor interest form if you are an instructor who is reviewing, using, or adapting *Significant Statistics* for a course, or if you have created any materials related to this book and would like to share them publicly.

Navigate to the book's main landing page to access:

- Links to multiple electronic versions of the textbook (PDF, EPUB, HTML)
- A link to order a print copy
- A link to the errata document
- A link to report errors

If you are interested in making your own version of this book, check out the book's license and information on recommended practices for attribution.

Helpful URLs

License: https://creativecommons.org/licenses/by-sa/4.0/

Significant Statistics website: https://sites.google.com/vt.edu/significantstatistics/home

Significant Statistics YouTube channel: https://www.youtube.com/channel/UCHVyc1NJuYvzpoom-L3nBpg

Significant Statistics podcast channel: https://anchor.fm/john-russell10

Instructor interest form: https://bit.ly/stat-interest

Book's main landing page: https://doi.org/10.21061/significantstatistics

View known errata: http://bit.ly/stat-errata

Report an error: https://bit.ly/feedback-stat

Recommended practices for attribution: https://wiki.creativecommons.org/wiki/Best_practices_for_attribution

CHAPTER 1: SAMPLING AND DATA

1.1 Introduction to Statistics

Learning Objectives

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

- Recognize and differentiate between key terms dealing with statistics
- Identify different types of data
- Identify data collection methods and study designs
- Apply various types of sampling methods to data collection

Introduction

We encounter statistics in our daily lives more often than we probably realize in many different contexts, such as news and weather reports or in lab and classroom settings.

"Statistics' ultimate goal is translating data into knowledge." – *Alan Agresti & Christine Franklin*

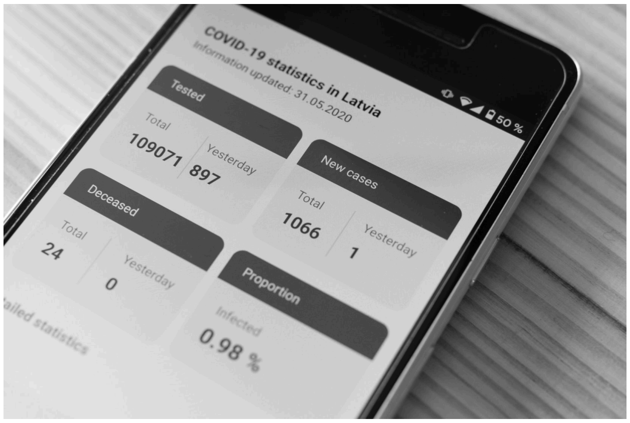


Figure 1.1: Smartphone display of COVID-19 statistics. Since the Coronavirus is novel, statisticians must collect and translate data into digestible information to give to the public. *[Figure description available at the end of the section](#)*.

You are probably asking yourself, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about everything from crime and politics to sports, education, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the best educated guess.

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Your chosen profession may very well involve some statistical knowledge. For example, the fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and terms of probability and statistics. You will soon understand how statistics and probability work together. You will also learn how data are gathered and how “good” data can be distinguished from “bad.”

The Study of Statistics

We see and use data in our everyday lives. The science of statistics deals with the collection, analysis, interpretation, and presentation of data. This is reflected in the **data analysis process**, which we will expand on in the next section.

You will first learn how to organize and summarize data. Organizing, summarizing, and presenting data is the basis of **descriptive statistics**. Data can be summarized with graphs or with numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing useful conclusions from data while filtering out the noise. These formal methods are called **inferential statistics**.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination. You will encounter a lot of mathematical formulas that seem to require calculations. Keep in mind, however, that the goal of statistics is not to perform numerous calculations using the formulas, but to interpret data to gain an understanding. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life. Statistical inference uses probability to determine how confident you can be that your conclusions are correct.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, a regular pattern emerges when there are many repetitions. After reading about the English statistician Karl Pearson tossing a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times, resulting in 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498, which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or of you getting an A in this course, we use probabilities. Doctors use probability to determine the chance of a medical test incorrectly diagnosing the presence of a disease. A stockbroker uses probability to determine the rate of return on a client’s investments. You might use probability to decide if you should buy a lottery ticket. In your study of statistics, you will utilize the power of mathematics and probability to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of people or things under study. To study the population, we select a sample. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. **Parameters** are numbers that describe a characteristic of the population.

Since it can take a great deal of resources (time, money, manpower, etc.) to examine an entire population, we often study only a subset of that population. Taking a **sample** is a very practical technique for accomplishing this. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion polls take samples of 1,000–2,000 people to represent the views of the entire country's population. Manufacturers of canned carbonated drinks take samples to determine if a 16-ounce can contains 16 ounces of carbonated drink.

From the information we collect in our sample, we can calculate a **statistic**. A statistic is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A parameter is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned by each student across all math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a representative sample. We are interested in both sample statistics and population parameters in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

Individuals are the units about which we are collecting information. This could be a person, animal, thing, or place. A **variable**, usually represented by capital letters such as X or Y, is a specific characteristic or measurement that can be determined for each individual. The **values** of a variable are the possible observations of the variable. If there are multiple variables collected on an individual, the entire set of variables may be called a case or observational unit.

Data refers to the actual values of the variables of interest. Data may be numbers or words. We'll dive into data in the next section.

Example

Determine how the key terms apply to the following study. We want to know the average (mean) amount of money first-year college students spend at ABC College on school supplies (excluding books). We randomly survey 100 first-year students at the college. Three of those students spent $150, $200, and $225.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here: *https://pressbooks.lib.vt.edu/significantstatistics/?p=23#h5p-2*

Your Turn!

Determine how the key terms apply to the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 1.1: Markus Winkler (2020). *Corona death and new cases stats*. Unsplash license. https://unsplash.com/photos/tUEnyweZjEU

Figure Descriptions

Figure 1.1: iPhone displaying COVID-19 statistics sits on a gray desk.

1.2 Data Basics

Types of Data

Data may come from a **population** or from a **sample**. Lowercase letters like x or y are generally used to represent data values. Most data falls into the following categories:

- **Qualitative (categorical)**
- **Quantitative (numerical)**

Qualitative or categorical data come in many forms. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Categorical data can generally be described with words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+.

Quantitative data, also known as numerical data, always takes the form of numbers. Quantitative data is typically the result of counting or measuring attributes of a population (e.g., amount of money, pulse rate, weight, number of people living in your town, or the number of students taking statistics). Quantitative data may be either **discrete** or **continuous**.

All data that results from counting is called quantitative discrete data. This data takes on only certain numerical values. If you count the number of phone calls you receive each day of the week, you might get values such as zero, one, two, or three.

Data that is composed not just of counted numbers but of all possible values on an interval (the real numbers) is called quantitative continuous data. Continuous data is often the result of measurements like lengths, weights, or durations. The length of a phone call in minutes would be quantitative continuous data.

Figure 1.2: Red Jaguar. Car type (in this case, Jaguar) can be considered categorical data since it is described using words. [Figure description available at the end of the section](#).

If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then it would fall into categories such as Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (for example, calculating the average points earned in class), but it makes no sense to do math with values of Y, as you can't calculate an average party affiliation.

Example

At the supermarket, you purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetables (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).

Name datasets that are quantitative discrete, quantitative continuous, and qualitative.

Possible solutions

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

Try to identify additional data sets in this example.

Your Turn!

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.lib.vt.edu/significantstatistics/?p=26#h5p-4

Levels of Measurement

The way a set of data is measured is called its level of measurement. The accuracy of statistical procedures depends on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement (from lowest to highest):

- **Nominal scale**
- **Ordinal scale**
- **Interval scale**
- **Ratio scale**

Data that is measured using a nominal scale is categorical data where the categories have no natural order. Colors, names, labels and favorite foods, as well as yes or no responses, are examples of nominal level data. For example, it's not possible to rank people according to their favorite food; putting pizza first and sushi second does not create meaningful data. Smartphone companies are another example of nominal scale data. The data are the companies that make smartphones, but there is no agreed-upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an ordinal scale is similar to nominal scale data, but there is a big difference. Ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five parks can be ranked from one to five, but we cannot measure differences between the data. Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are "excellent," "good," "satisfactory," and "unsatisfactory." These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be quantified. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the interval scale is similar to ordinal level data because it has a definite order. However, there is a meaningful difference between values of the data from an arbitrary starting point. Temperature scales like Celsius (C) and Fahrenheit (F) are measured using the interval scale. In both temperature measurements, differences make sense, but 40° is equal to 100° minus 60°. But 0°F and 0°C do not align because 0 is not the absolute lowest temperature in both scales. Temperatures like -10°F and -15°C exist and are colder than 0. Interval level data can be used in calculations, but one type of comparison cannot be made. 80°C is not four times as hot as 20°C (nor is 80°F four times as hot as 20°F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the ratio scale takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20, and 92 (out of a possible 100 points). The exams are machine graded. The data can be put in order from lowest to highest: 20, 68, 80, 92. The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

NOTE: You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Your Turn!

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.lib.vt.edu/significantstatistics/?p=26#h5p-5

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16-ounce cans were measured and produced the following amount of beverage (in ounces): 15.8, 16.1, 15.2, 14.8, 15.8, 15.9, 16.0, 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people took the measurements or because the exact amount (16 ounces of liquid) was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range. As you take data, be aware that yours may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to re-evaluate your data-taking methods and your accuracy.

Data Analysis

In this age of “Big Data,” **data analysis** is an essential tool. Informally, it could be defined as the process of collecting, organizing, and analyzing your data. Formally, the process consists of four phases with associated questions:

1. **Identify the research objective.**
    - What questions are to be answered?
    - What group should be studied?
    - Have attempts been made to answer it before?
2. **Collect the information needed.**
    - Is data already available?
    - Can you access the entire population?
    - How can you collect a good sample?
3. **Organize and summarize the information.**
    - What visual descriptive techniques are appropriate?
    - What numerical descriptive techniques are appropriate?
    - What aspects of the data stick out?
4. **Draw conclusions from the information.**
    - What inferential techniques are appropriate?
    - What conclusions can be drawn?

We will answer all of these questions and more throughout the course.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 1.2: Mateusz Delegacz (2017). *London Jaguar 2*. Unsplash license. https://unsplash.com/photos/1Ah8CAwk3vM

Figure Descriptions

Figure 1.2: A red Jaguar car sits on the street in front of a building.

1.3 Data Collection and Observational Studies

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? When we are interested in the effect one variable may have on another, we call the first variable the **explanatory variable** and the second the **response variable.** Questions like these are answered using studies and experiments. Proper study design ensures the production of reliable, accurate data.

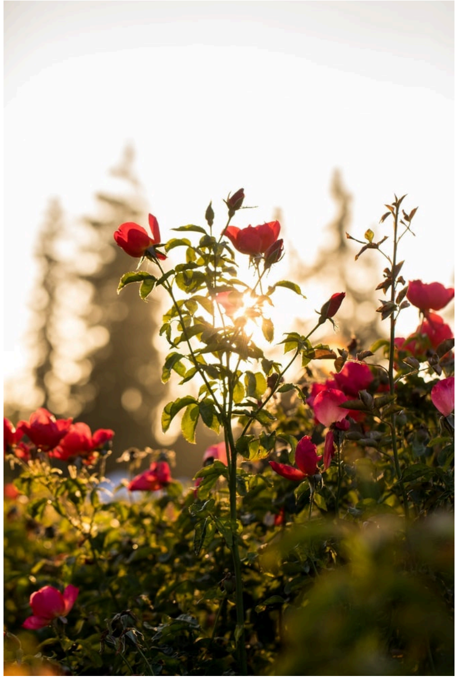


Figure 1.3: Flower growth. Is one brand of fertilizer more effective at growing flowers than another? Statisticians can answer this question by determining what effect the explanatory variable (fertilizer brands) has on the response variable (flower growth). Figure description available at the end of the section.

Data Collection Methods

There are many ways **data** is commonly collected, each with their own advantages and disadvantages. Some ways data may be collected are:

- **Anecdotal evidence**
- **Observational studies**
- **Designed (controlled) experiments**

The latter two options are more commonly accepted, but we will briefly describe the former first.

Anecdotal Evidence

Consider the following statements seemingly based on data:

1. I met two students who took more than seven years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
2. A man on the news had an adverse reaction to a vaccine, so it must be dangerous.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Though each conclusion is technically based on data, there are two problems. First, the data in each example only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion is called anecdotal evidence.

While such evidence may be true and verifiable, be careful of data collected in this way since it may only represent extraordinary or unusual cases. Often, we are more likely to recall anecdotal evidence based on its striking characteristics. For instance, in Case #1 above, we are more likely to remember the two people we met who took seven years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

Observational Studies

Researchers perform an observational study when they collect data in a way that does not directly interfere with how the data arises. For instance, researchers may collect information via a questionnaire or survey, review medical or company records, or follow a large group of similar individuals to form hypotheses about why certain diseases develop. In each of these situations, researchers merely observe the data that occurs. In general, observational studies can provide evidence of naturally occurring **associations** between variables, but they cannot by themselves show a causal connection. Why not? Consider the following example.

Suppose an observational study tracking sunscreen use and skin cancer found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer? Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent may be sun exposure.

Exposure to the sun is unaccounted for in this simple investigation, even though it stands to reason if someone is out in the sun all day, she is more likely to use sunscreen *but also* more likely to get skin cancer. Sun exposure here is an example of what we might call a **confounding variable**. Also known as a lurking or conditional variable, this is a variable that was not accounted for and may actually be important. Confounding variables can cause many misleading, counterintuitive, or even humorous (spurious) correlations.

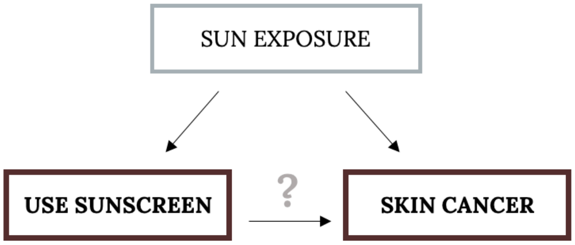


Figure 1.4: Association between sunscreen and skin cancer. Figure description available at the end of the section.

Observational studies come in two forms: **prospective** and **retrospective**. A prospective study identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of patients over many years to assess the possible influences of behavior on cancer risk. One example of such a study is the Nurses' Health Study, started in 1976 and expanded in 1989. This prospective study recruits registered nurses and then collects data from them using questionnaires. Retrospective studies collect data after events have taken place (e.g., researchers reviewing past events in medical records). Some datasets may contain both prospectively and retrospectively collected variables.

There are other classifications of observational studies you may encounter, especially in life science and medical contexts. A **cohort study** follows a group of many similar individuals over time, often producing **longitudinal** data. A **cross-sectional study** indicates data collection on a population at one point in time (often prospective). A **case-control study** compares a group that has a certain characteristic to a group that does not, often taking the form of a retrospective study for rare conditions.

Example

A researcher is studying the relationship between time spent studying in medical school and depression rates among students. The researcher looks at graduated students' medical records to determine if they have ever seen a psychologist. He also sends out a questionnaire to the same students to ask how much time they spent studying in college. What type of study is this?

Solution
This is both a prospective and retrospective observational study. Sending out a questionnaire indicates a prospective study, while reviewing past medical records indicates a retrospective study.

Your Turn!

A researcher is wondering if the same individual can contract COVID-19 more than once. She randomly selects 300 people who have tested positive for COVID-19. The participants fill out a self-report survey once a month to inform the researcher if they have tested positive again. What type of study is this?

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 1.3: Jason Leung (2018). *Selective focus photo of red peonies*. Unsplash license. https://unsplash.com/photos/nonlZlChSZQ

Figure 1.4: Kindred Grey (2020). *Association between sunscreen and skin cancer.* CC BY-SA 4.0.

Figure Descriptions

Figure 1.3: Close up photo of a peony bush at sunrise.

Figure 1.4: Three boxes form a triangle. The top box reads 'sun exposure' and has arrows pointing to two boxes. The bottom left box reads 'use sunscreen' and the bottom right reads 'skin cancer'. There is an arrow with a question mark pointing from the bottom left box to the bottom right one.

1.4 Designed Experiments

Observational Studies vs. Experiments

Ignoring anecdotal evidence, there are two primary types of data collection: **observational studies** and **controlled (designed) experiments**. Remember, we typically cannot make claims of causality from observation studies because of the potential presence of confounding factors. However, making causal conclusions based on experiments is often reasonable if we control for those factors.

Suppose you want to investigate the effectiveness of vitamin D in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin D. You notice that the subjects who take vitamin D exhibit better health on average than those who do not. Does this prove that vitamin D is effective in disease prevention? It does not. There are many differences between the two groups beyond just vitamin D consumption. People who take vitamin D regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could influence health. As described, this study does not necessarily prove that vitamin D is the key to disease prevention.

Experiments ultimately aim to provide evidence for use in decision-making, so how could we narrow our focus and make claims of causality? In this section, you will learn important aspects of experimental design.

Designed Experiments

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable may be called **treatments**. An **experimental unit** is a single object or individual being measured.

The main principles to follow in experimental design are:

1. Randomization
2. Replication
3. Control

Randomization

In order to provide evidence that the explanatory variable is indeed causing the changes in the response variable, it is necessary to isolate the explanatory variable. The researcher must design the experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by randomizing the experimental units placed into treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point, the only difference between groups is the one imposed by the researcher. As a result, different outcomes measured in the response variable must be a direct result of the different treatments. In this way, an experiment can show an apparent cause-and-effect connection between the explanatory and response variables.

Recall our previous example of investigating the effectiveness of vitamin D in preventing disease. Individuals in our trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: the control group (no treatment) and the experimental group (extra doses of vitamin D).

Replication

The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we replicate by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding. It is also helpful to subject individuals to the same treatment more than once, which is known as **repeated measures**.

Control

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectations of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted, "*Results showed that believing one had taken the substance resulted in* [performance] *times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*" [1]

It is often difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment—a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill that contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind** experiment is one in which both the subjects and the researchers involved with the subjects are unaware.

Randomized experiments are an essential tool in research. The U.S. Food and Drug Administration typically requires that a new drug can only be marketed after two independently conducted randomized trials confirm its safety and efficacy; the European Medicines Agency has a similar policy. Large randomized experiments in medicine have provided the basis for major public health initiatives. In 1954, approximately 750,000 children participated in a randomized study comparing the polio vaccine with a placebo. In the United States, the results of the study quickly led to the widespread and successful use of the vaccine for polio prevention.

Example

How does sleep deprivation affect your ability to drive? A recent study measured its effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=33#h5p-7

Your Turn!

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed pencil-and-paper mazes multiple times while wearing masks. They completed the mazes three times wearing floral-scented masks and three times with unscented masks. Participants were assigned at random to wear the floral mask during either the first three or last three trials. For each trial, researchers recorded the time it took to complete the maze and whether the subject's impression of the mask's scent was positive, negative, or neutral.

1. Describe the explanatory and response variables in this study.
2. What are the treatments?
3. Identify any lurking variables that could interfere with this study.
4. Is it possible to use blinding in this study?

Solution

1. The explanatory variable is scent and the response variable is the time it takes to complete the maze.
2. There are two treatments: a floral-scented mask and an unscented mask.
3. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
4. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded. The researcher who is observing a subject will not know which mask is being worn.

More Experimental Design

There are many different experimental designs from the most basic—a single treatment and control group—to some very complicated designs. When working with more than one treatment in an experimental design setting, these variables are often called **factors**, especially if they are categorical. The values of factors are are often called **levels**. When there are multiple factors, the combinations of each of the levels are called **treatment combinations**, or interactions. Some basic types of interactions you may see are:

1. **Completely randomized**
2. **Block design**
3. **Matched pairs design**

Completely Randomized

This essential research tool does not require much explanation. It involves figuring out how many treatments will be administered and randomly assigning participants to their respective groups.

Block Design

Researchers sometimes know or suspect that variables outside of the treatment influence the response. Based on this, they may first group individuals into blocks and then randomly draw cases from each block for the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in the figure below. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

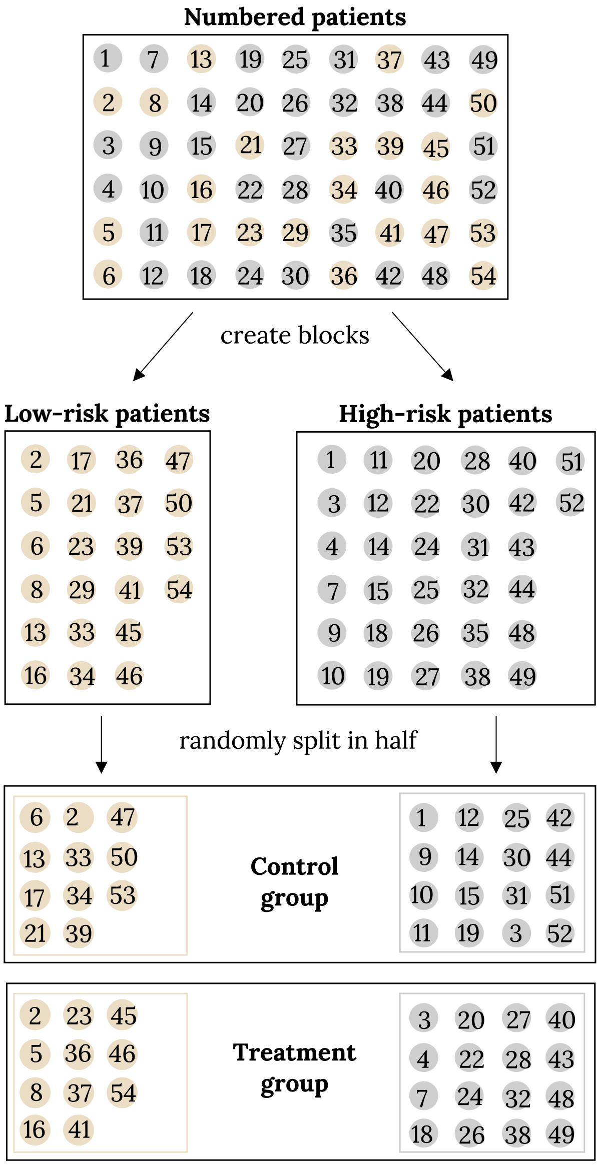


Figure 1.5: Block design. Figure description available at the end of the section.

Matched Pairs

A **matched pairs design** is one where very similar individuals (or even the same individual) receive two different treatments (or treatment vs. control) and the results are compared. Though this design is very effective, it can be hard to find many suitably similar individuals. Some common forms of a matched pairs design are twin studies, before-and-after measurements, pre- and post-test situations, and crossover studies.

Was the use of a new wetsuit design responsible for an observed increase in swim velocities at the 2000 Summer Olympics? In a matched pairs study designed to investigate this question, twelve competitive swimmers swam 1,500 meters at maximal speed, once wearing a wetsuit and once wearing a regular swimsuit. The order of wetsuit and swimsuit trials was randomized for each of the 12 swimmers. Figure 1.6 shows the average velocity recorded for each swimmer, measured in meters per second (m/s).

	swimmer.number	wet.suit.velocity	swim.suit.velocity	velocity.diff
1	1	1.57	1.49	0.08
2	2	1.47	1.37	0.10
3	3	1.42	1.35	0.07
4	4	1.35	1.27	0.08
5	5	1.22	1.12	0.10
6	6	1.75	1.64	0.11
7	7	1.64	1.59	0.05
8	8	1.57	1.52	0.05
9	9	1.56	1.50	0.06
10	10	1.53	1.45	0.08
11	11	1.49	1.44	0.05
12	12	1.51	1.41	0.10

Figure 1.6: Average velocity of swimmers

In this data, two sets of observations are uniquely paired so that an observation in one set matches an observation in the other; in this case, each swimmer has two measured velocities, one with a wetsuit and one with a swimsuit. A natural measure of the effect of the wetsuit on swim velocity is the difference between the measured maximum velocities (velocity.diff = wet.suit.velocit − swim.suit.velocity). Even though there are two measurements per individual, using the difference in observations as the variable of interest allows for the problem to be analyzed.

Example

A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. What experiment design is being implemented here?

Solution
Matched pairs

Your Turn!

A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The mean hours slept for each person were recorded before and after stating the medication. What experiment design is being implemented here?

Solution
Matched pairs

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 1.5: Kindred Grey (2020). *Block design.* CC BY-SA 4.0.

Figure Descriptions

Figure 1.5: Box labeled 'numbered patients' that has 54 blue or orange circles numbered from one through 54. Two arrows point from this box to two boxes below it with the caption 'create blocks'. The left box is all of the orange circles grouped together labeled 'low-risk patients'. The right box is all of the blue circles grouped together labeled 'high-risk patients'. An arrow points down from the left box and the right box with the caption 'randomly split in half'. The arrows point to a 'Control' box and a 'Treatment' box. Both of these boxes have half orange circles and half blue circles.

Notes

1. McClung, Mary, and Dave Collins, ""Because I know it will!": Placebo Effects of an Ergogenic Acid on Athletic Performance," *Journal of Sport & Exercise Psychology*, 29, no. 3 (2007): 382-394.

1.5 Sampling

Sampling

Gathering information about an entire population is often virtually impossible due to costs or other factors. Instead, we typically use a **sample** of the population which should have the same characteristics as the population it is representing. Statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of random sampling. In each form, each member of a population initially has an equal chance of being selected for the sample. There are advantages and disadvantages to each sampling method.

Simple Random Sampling

The gold standard and maybe easiest method to describe is called a **simple random sample (SRS)**. Any group of n individuals is equally likely to be chosen as any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected.

For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names.

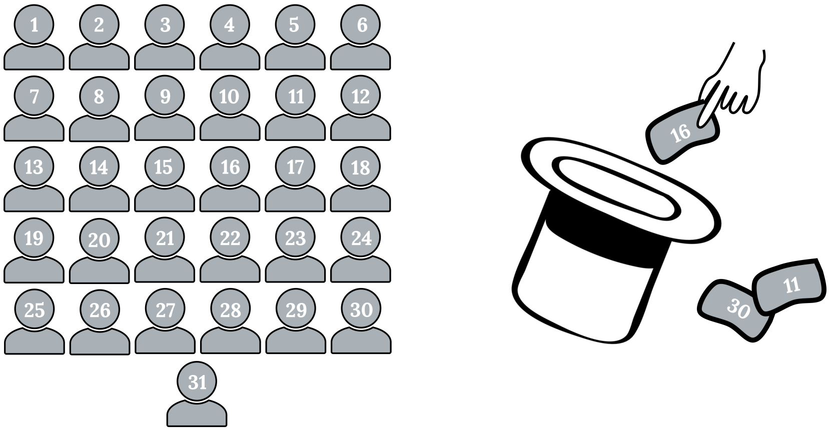


Figure 1.7: Simple random sample. Figure description available at the end of the section.

A more technological approach is for Lisa to pair the last name of each class member with a two-digit number, as in the table below:

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Figure 1.8: Lisa's class roster

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa uses a calculator to generate the following random numbers:

0.94360, 0.99832, 0.14669, 0.51470, 0.40581, 0.73381, 0.04399

Lisa identifies multiple two-digit numbers in each of these random numbers (i.e., 0.94360 becomes 94, 43, 36, and 60). If any of these two-digit numbers corresponds with a name on her list, that student is chosen. She can generate more random numbers if necessary.

The random numbers 0.94360 and 0.99832 do not contain appropriate two-digit numbers. However, the third random number, 0.14669, contains 14, the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

Other Sampling Techniques

In addition to simple random sampling, there are other forms of sampling that involve a chance process in getting the sample. Other well-known random sampling methods are:

- **Stratified sampling**
- **Cluster sampling**

- **Systematic sampling**

To choose a **stratified sample**, identify a relevant similar characteristic of your population and divide people into groups, or strata, based on this characteristic. Then take a proportionate number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each department to get a stratified random sample. Note that there are six individuals in department one, 12 in department two, and nine in department three. If we wanted a total sample size of nine with equal representation from each department, we would randomly choose two individuals from department one, four from department two, and three from department three to make up the sample. Stratified random sampling is often used when we want to make sure our sample is representative of population demographics.

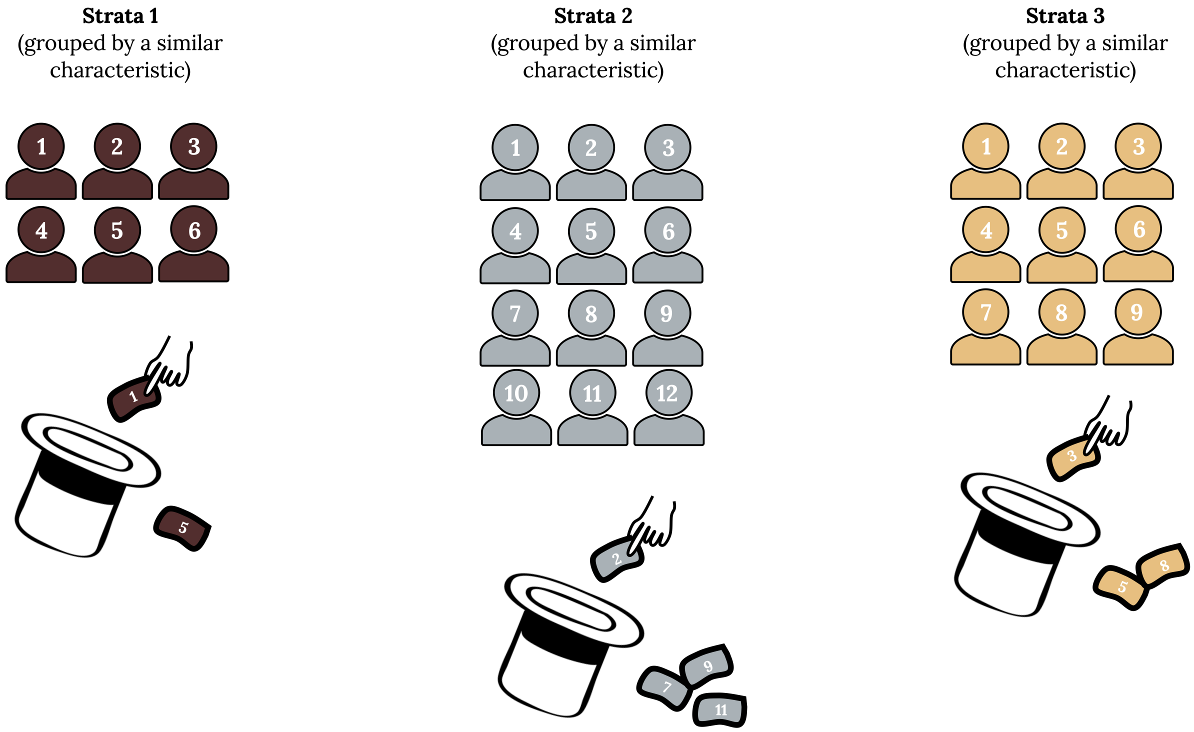

Strata 1
(grouped by a similar characteristic)
1
2
3
4
5
6
1
5
Strata 2
(grouped by a similar characteristic)
1
2
3
4
5
6
7
8
9
10
11
12
2
9
7
11
Strata 3
(grouped by a similar characteristic)
1
2
3
4
5
6
7
8
9
3
8
5

Figure 1.9: Stratified sample. Figure description available at the end of the section.

To choose a **cluster sample**, the population will need to be divided into predefined clusters or groups. Then, randomly select some of the clusters. All of the members from the selected clusters now make up your sample. For example, suppose your college has 5 departments pictured in different colors below. Each of these departments are the clusters. Number each department and choose two different numbers using simple random sampling. All members of the two chosen departments (grey and green) are the cluster sample.

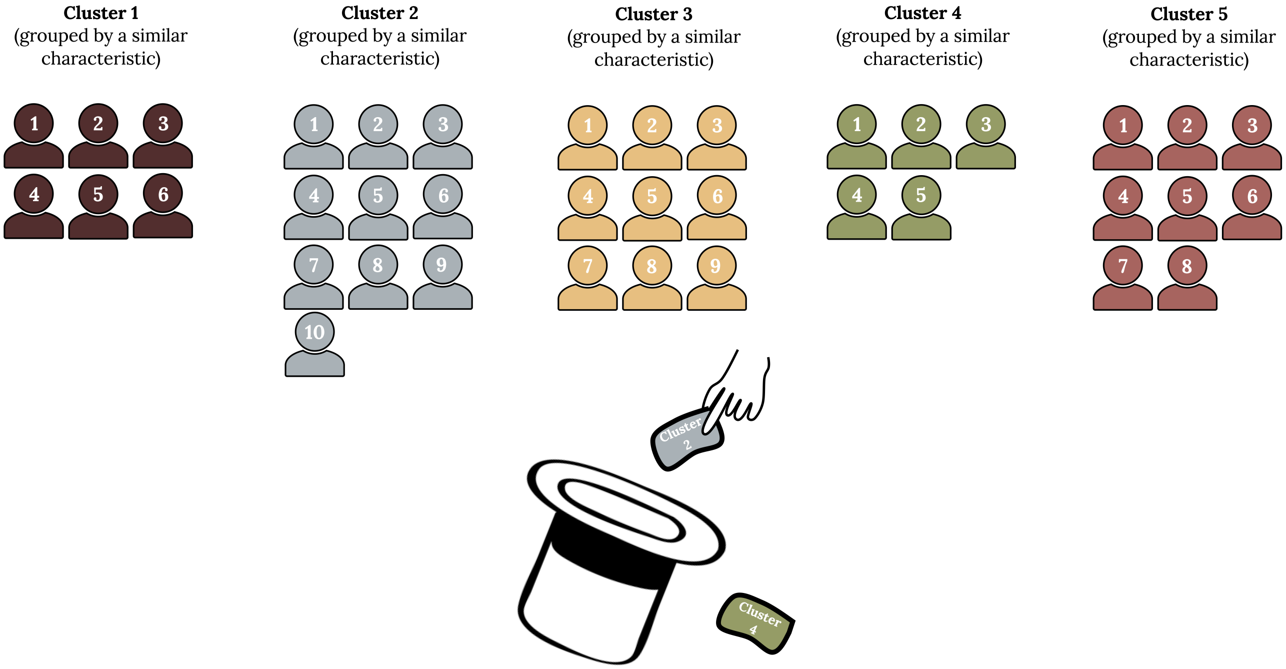

Cluster 1
(grouped by a similar characteristic)
1 2 3 4 5 6
Cluster 2
(grouped by a similar characteristic)
1 2 3 4 5 6 7 8 9 10
Cluster 3
(grouped by a similar characteristic)
1 2 3 4 5 6 7 8 9
Cluster 4
(grouped by a similar characteristic)
1 2 3 4 5
Cluster 5
(grouped by a similar characteristic)
1 2 3 4 5 6 7 8
Cluster 2
Cluster 4

Figure 1.10: Cluster sample. *Figure description available at the end of the section.*

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. You have 60 contacts in your phone but can't call all of them, so you decide on a sample size of 15. Number the population 1–60 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fifth name thereafter until you have a total of 15 names. Systematic sampling is frequently chosen because it is a simple method.

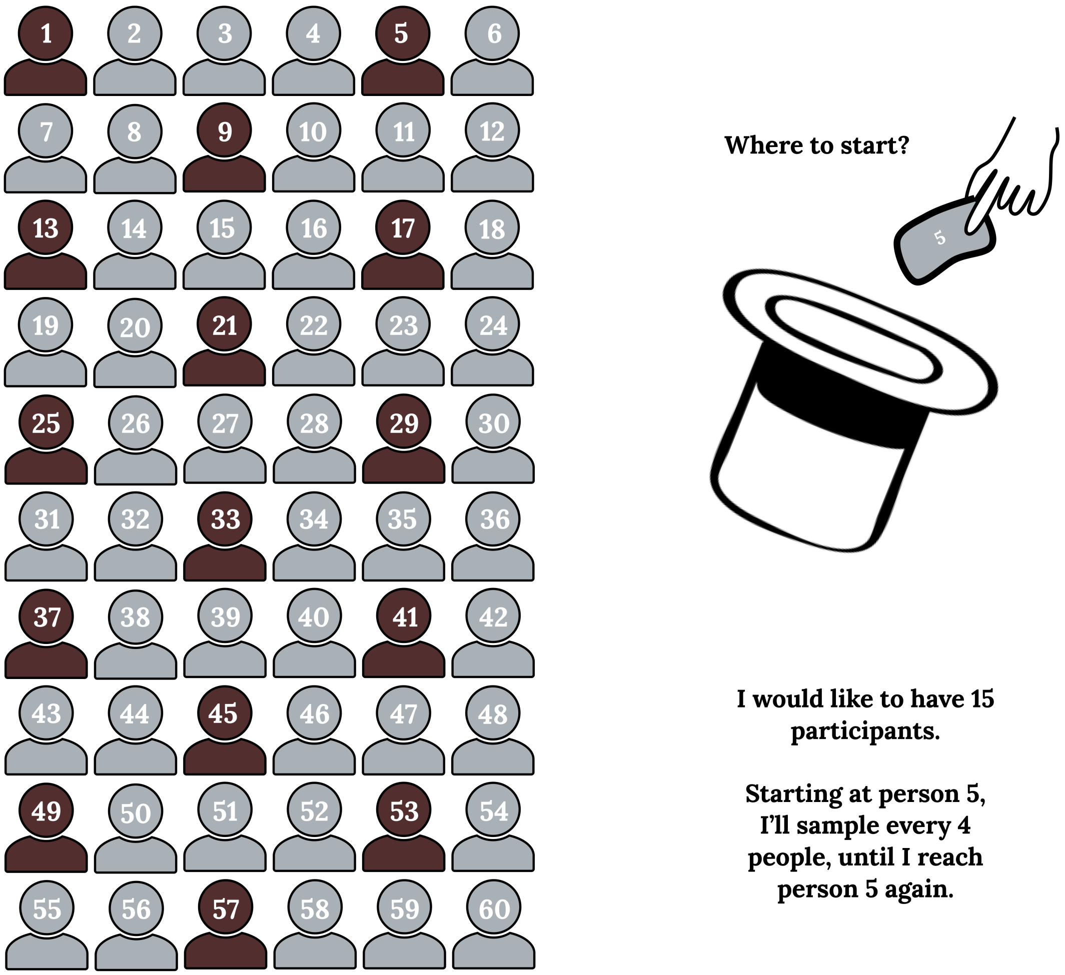


Figure 1.11: Systematic sample. Figure description available at the end of the section.

Researchers may also choose to some a combination of these techniques, called multistage sampling.

Example

A study is conducted to determine the average tuition that Virginia Tech undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the fall semester. What is the type of sampling in each case?

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here: *https://pressbooks.lib.vt.edu/significantstatistics/?p=35#h5p-10*

Your Turn!

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task. The station uses convenience sampling and surveys the first 200 people they meet at one of the station's concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music. Do you think that this sample is representative of (or characteristic of) the entire 20,000 listener population?

Sampling and Replacement

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However, for practical reasons, simple random sampling is done in most populations **without replacement**. Surveys are typically done without replacement, where a member of the population may be chosen only once. Most samples are taken from large populations, and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. For any particular sample of 1,000, if you are sampling with replacement,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 10,000 (0.0999);

- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling without replacement,

- the chance of picking the first person for any particular sample is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions $\frac{999}{10,000}$ and $\frac{999}{9,999}$. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling with replacement for any particular sample, then the chance of picking the first person is 10 out of 25, and the chance of picking a different second person is 9 out of 25 (you replace the first person).

If you sample without replacement, then the chance of picking the first person is 10 out of 25, and then the chance of picking the second (different) person is 9 out of 24 (you do not replace the first person).

Compare the fractions $\frac{9}{25}$ and $\frac{9}{24}$. To four decimal places, $\frac{9}{25}$ = 0.3600 and $\frac{9}{24}$ = 0.3750. To four decimal places, these numbers are not equivalent.

Bias in Samples

Sampling data should be done very carefully and collecting data carelessly can have devastating results. For example, surveys mailed to households and then returned may be very biased (they may favor a certain group). It is often best for the person conducting the survey to select the sample respondents.

When you analyze data, it is important to be aware of sampling errors and non-sampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause non-sampling errors. A defective counting device can cause a non-sampling error.

In reality, a sample will never be exactly representative of the population, so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others. Remember, each member of the population should have an equally likely chance of being chosen. When sampling bias occurs, incorrect conclusions can be drawn about the population being studied.

Variation in Samples

As previously mentioned, two or more samples from the same population, taken randomly and having close to the same characteristics of the population, will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling, and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, the samples would surely be different. Neither would be wrong, however.

Think about why Doreen's and Jung's samples would be different.

If Doreen and Jung took larger samples (i.e., the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. Even then, their samples would be, in all likelihood, different from each other. This idea of **sampling variability** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a couple hundred observations (or even fewer) are sufficient for many purposes. In polling, samples consisting of 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and well done. You will learn why when you study confidence intervals.

Be aware that even many large samples are biased. For example, call-in surveys are invariably biased because people choose to respond or not.

Critical Evaluation

We need to evaluate statistical studies critically and analyze them before accepting their results. Common problems include:

- **Convenience sampling:** A type of sampling that is non-random and involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favoring certain outcomes) in others.
- **Problems with samples:** A sample must be representative of the population. A sample that is not repre-

sentative of the population is biased. Biased, unrepresentative samples give results that are inaccurate and not valid.

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

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that “numbers don’t lie,” but the people who use numbers to support their claims often do.

A recent investigation into famous social psychologist Diederik Stapel has led to the retraction of his articles from some of the world’s top journals, including *Journal of Experimental Social Psychology*, *Social Psychology*, *Basic and Applied Social Psychology*, and the *British Journal of Social Psychology*, as well as the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and ten PhD dissertations that he supervised.

> *Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. “It was a quest for aesthetics, for beauty—instead of the truth,” he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.*[1]

The committee investigating Stapel found him guilty of several misdeeds, including creating datasets that largely confirmed prior expectations, altering data in existing datasets, changing measuring instruments without reporting the change, and misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[2] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations like the American Statistical Association clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers are mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as Institutional Review Boards (IRBs). All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give informed consent. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and unanticipated risks arise? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point, the blood becomes biological waste. Does a researcher have the right to use it in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (http://www.retractionwatch.com/) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Example

A researcher is collecting data in a community. Describe the unethical behavior in each example and how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

1. She selects a block where she is comfortable walking because she knows many of the people living on the street.
2. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
3. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution

1. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
2. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
3. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

Your Turn!

Describe the unethical behavior, if any, in each example and how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

a. The survey is commissioned by the seller of a popular brand of apple juice.
b. There are only two types of juice included in the study: apple juice and cranberry juice.
c. Researchers allow participants to see the brand of juice as each sample is poured for a taste test.
d. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y, and 42% have no preference between the two brands. Brand X references the study in a commercial, saying "Most teens like Brand X as much as or more than Brand Y.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 1.7: Kindred Grey (2024). *Simple random sample*. CC BY-SA 4.0.

Figure 1.9: Kindred Grey (2024). *Stratified sample*. CC BY-SA 4.0.

Figure 1.10: Kindred Grey (2024). *Cluster sample*. CC BY-SA 4.0.

Figure 1.11: Kindred Grey (2024). *Systematic sample*. CC BY-SA 4.0.

Figure Descriptions

Figure 1.7: 31 people. Put numbers one through 31 into a hat and draw out three numbers.

Figure 1.9: Three groups. All grouped by similar characteristics. Group one has six people. Put six numbers in a hat and draw two. Group two has 12 people. Put 12 numbers in a hat and draw four. Group three has nine people. Put nine numbers in a hat and draw three.

Figure 1.10: Five groups with varying numbers of people all grouped by a similar characteristic. Put five numbers in a hat and draw two. Use all the people in both of these groups as your sample size.

Figure 1.11: 60 people. Put 60 numbers in a hat and pull out one random number. That is the first person you'll sample. If you want to have a sample size of 15, then sample every four people until you reach the number you chose from the hat. It's okay to loop back to the first person.

Notes

1. Yudhijit Bhattacharjee, "The Mind of a Con Man," Magazine, *New York Times*, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).
2. "Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," Tillburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd-bce5-4385-b9ff-05b840caeae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).

Chapter 1 Wrap-Up

Concept Check

Take this quiz to check your comprehension of this chapter.

If you are using an offline version of this text, access the quiz for this chapter via the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=37#h5p-12

Section Resources

If you are using an offline version of this text, access these materials by visiting https://doi.org/10.7294/26207456.

1.1 Introduction to Statistics and Key Terms

1.2 Data Basics

1.3 Data Collection and Observational Studies

1.4 Designed Experiments

1.5 Sampling

Key Terms

Try to define the terms below on your own. Check your response by clicking on the term, or looking at the end-of-book glossary!

1.1 Introduction to Statistics and Key Terms

- **Data analysis process**
- **Descriptive statistics**
- **Inferential statistics**
- **Probability**
- **Population**
- **Parameters**
- **Sample**
- **Statistic**
- **Individuals**
- **Variable**
- **Values**
- **Data**

1.2 Data Basics

- **Qualitative (categorical)**
- **Quantitative (numerical)**
- **Discrete**
- **Continuous**
- **Nominal scale**
- **Ordinal scale**
- **Interval scale**
- **Ratio scale**
- **Variation**
- **Data analysis**

1.3 Data Collection and Observational Studies

- **Explanatory variable**
- **Response variable**
- **Anecdotal evidence**
- **Observational studies**
- **Designed (controlled) experiments**
- **Associations**
- **Confounding (lurking, conditional) variable**
- **Prospective study**

- **Retrospective study**
- **Cohort study**
- **Longitudinal study**
- **Cross-sectional study**
- **Case-control study**

1.4 Designed Experiments

- **Treatments**
- **Experimental unit**
- **Repeated measures**
- **Control group**
- **Placebo**
- **Blinding**
- **Double-blind**
- **Factors**
- **Levels**
- **Treatment combinations (interactions)**
- **Completely randomized**
- **Block design**
- **Matched pairs design**

1.5 Sampling Techniques and Ethics

- **Simple random sample (SRS)**
- **Stratified sampling**
- **Cluster sampling**
- **Systematic sampling**
- **Sampling bias**
- **Sampling variability**
- **Convenience sampling**

Extra Practice

Extra practice problems are available at the end of the book (Chapter 1 Extra Practice).

CHAPTER 2: UNIVARIATE DESCRIPTIVE STATISTICS

2.1 Descriptive Statistics and Frequency Distributions

Learning Objectives

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

- Display and interpret categorical data
- Display and interpret quantitative data
- Recognize, describe, and calculate the measures of the center of quantitative data
- Recognize, describe, and calculate the measures of spread of quantitative data
- Recognize, describe, and calculate the measures of location of quantitative data
- Identify outliers in quantitative data

Descriptive Statistics

Once you collect data, what do you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. If you have no clue about house prices, you might ask your real estate agent to give you a sample dataset of prices, but looking through all the prices can be overwhelming. A better way might be to look at numerical descriptions such as the average or median house price. Your agent might also provide you with a graph of the data.

Figure 2.1: Voting ballots. When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. Figure description available at the end of the section.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **descriptive statistics.** We will look at both **graphical** and **numerical** descriptive methods. You will learn how to construct, calculate, and, most importantly, interpret these measurements and graphs.

Numerical descriptors consist of summary statistics (typically calculated from a sample) that represent important aspects such as the central tendency and variability of a distribution or the relative standing of a single observation with regard to the rest of the distribution.

Graphical descriptive methods consist of chart, tables, and graphs. These are tools that help you learn about the **distribution,** or shape, of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where the data clusters and where there are only a few data values. Newspapers and the internet sources use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data before more formal tools are applied.

The type of graph you choose first depends on the type of data with which you are working. Some of the types of graphs used to display categorical data are pie charts and bar charts. Some graphs that are used to summarize and organize quantitative data are the dot plot, the histogram, the stem-and-leaf plot, the frequency polygon, the box plot, and, in special cases, the time series plot. The emphasis here will be on histograms and box plots.

We will start by looking at a graphical method that can display any type of data, the frequency table.

Frequency Tables

Frequency tables are a great starting place for summarizing and organizing your data. Once you have a set of data, you may first want to organize it to see the **frequency** (how often each value occurs in the set).

Frequency tables can be used to show either quantitative or categorical data. Displaying categorical data in a frequency table is fairly straightforward since you already have clearly defined categories. For example, if you polled 20 kindergarteners on their favorite colors, you could construct the following simple frequency table:

Color	Frequency
Red	2
Orange	2
Yellow	1
Green	3
Blue	4
Purple	3
Pink	4
Clear with sparkles	1
	Total = 20

Figure 2.2: Frequency table of children's favorite colors

Some quantitative data, especially discrete, may only a contain a limited number of values and little thought would be needed in creating the frequency table. Some data may have a natural grouping. For example, if you were organizing adults aged 20-69, it might make intuitive sense to group them as follows:

- 20–29
- 30–39
- 40–49
- 50–59
- 60–69

Consider the 30-39 grouping. Each group is typically called a class, or bin. In this case, 30 would be the **lower class limit**, while 39 is the **upper class limit.** The **class width** is defined as the difference between consecutive lower class limits. For the class 30–39, the class width is 40–30 = 10. The **class midpoint** is found by adding the lower limit and upper limit, then dividing by 2. For the class 30–39, the class midpoint would be calculated as follows:

$$\frac{30+39}{2} = 34.5$$

Depending on the format and precision of the data reported, we may have to decide how best to bin, or group, our data. Grouping data may not always be a clean or intuitive process. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05), which is more precise. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 – 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 – 0.0005 = 0.9995). If the data is entirely made up of integers and the smallest value is 2, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). When the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

The next question may concern how many bins to use. Generally anywhere from 5-20 bins, since too few does not display distribution well, but too many can have strange effects. A good place to start is the square root of your number of observations (*n*). Some other basic guidelines are that bins should not overlap or have gaps between them and should have the same width and cover the entire range of the data. The class limits and width should be "reasonable" numbers (e.g., whole numbers, or multiples of five or ten). In the end, it really just depends on the format of your data, but following these general guidelines should make sure your table is useful.

Relative Frequencies

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample—in this case, 20. Relative frequencies can be written as fractions, percents, or decimals. To find the relative frequency:

$$\text{RF} = \frac{f}{n}$$

in which:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies)
- *RF* = relative frequency

For example, if three students in Mr. Ahab's English class of 40 students received scores from 90% to 100%, then, $f = 3$, $n = 40$, and $\text{RF} = \frac{f}{n} = \frac{3}{40} = 0.075$. Of the students, 7.5% received scores between 90% and 100%. In this case, 90–100% are quantitative measures.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in the figure below.

NOTES:

- The sum of all frequencies will add up to n, or your sample size.
- All relative frequencies should add up to one (pending rounding).
- The first entry of the cumulative relative frequency column will be the same as the first entry of the relative frequency column since there is nothing to accumulate.
- The last entry of the cumulative relative frequency column is one, indicating that 100% of the data has been accumulated.

Example

The following table represents one way of grouping the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Heights (inches)	Frequency	Relative frequency	Cumulative relative frequency
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	0.05 + 0.03 = 0.08
63.95–65.95	15	$\frac{15}{100} = 0.15$	0.08 + 0.15 = 0.23
65.95–67.95	40	$\frac{40}{100} = 0.40$	0.23 + 0.40 = 0.63
67.95–69.95	17	$\frac{17}{100} = 0.17$	0.63 + 0.17 = 0.80
69.95–71.95	12	$\frac{12}{100} = 0.12$	0.80 + 0.12 = 0.92
71.95–73.95	7	$\frac{7}{100} = 0.07$	0.92 + 0.07 = 0.99
73.95–75.95	1	$\frac{1}{100} = 0.01$	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Figure 2.3: Frequency table of soccer player height

In this sample, there are five players whose heights fall within the interval 59.95–61.95 inches, three players whose heights fall within the interval 61.95–63.95 inches, 15 players whose heights fall within the interval 63.95–65.95 inches, 40 players whose height falls within the interval 65.95–67.95 inches, 17 players whose heights fall within the interval 67.95–69.95 inches, 12 players whose heights fall within the interval 69.95–71.95, seven players whose heights fall within the interval 71.95–73.95, and one player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

From the figure above, find the percentage of heights that are less than 65.95 inches.

Solution
If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then 23/100 or 23%. This percentage is the cumulative relative frequency entry in the third row.

Find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution
Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Use the heights of the 100 male semiprofessional soccer players. Fill in the blanks and check your answers.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.lib.vt.edu/significantstatistics/?p=41#h5p-41

What kind of data are the heights?

Solution
Quantitative continuous

Describe how you could gather this data (the heights) to make it characteristic of all male semiprofessional soccer players.

Solution
Get rosters from each team and choose a simple random sample from each.

Remember: you count frequencies. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Your Turn!

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, and 3. Construct a bar graph that shows the registered voter population by district.

Construct an appropriate table including frequencies, relative frequencies, and cumulative relative frequencies.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 2.1: U.S. Marine Corps photo by Staff Sgt. William Greeson (2009). US *Navy* 090821-M-0440G-043 *Voting ballots organized and arranged for counting by Afghan presidential election workers at a local school in the Nawa District*. Public domain. https://commons.wikimedia.org/wiki/File:US_Navy_090821-M-0440G-043_Voting_ballots_organized_and_arranged_for_counting_by_Afghan_presidential_election_workers_at_a_local_school_in_the_Nawa_District.jpg.

Figure Descriptions

Figure 2.1: This photo shows about 26 rolls of paper piled together. The rolls are different sizes.

2.2 Displaying and Describing Categorical Distributions

Descriptive Statistics for Categorical Data

Working with **categorical data** is typically more straightforward. Recall that descriptive statistics consists of visual and numerical methods. We usually start with visual methods and then move on to numerical.

Graphical Methods for Categorical Data

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make it easier to compare the same categories across colleges. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College				**Foothill College**		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Figure 2.4: Full- and part-time students

Tables are a good way of organizing and displaying data, but graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display categorical data are pie charts and bar graphs.

Pie Charts

In a pie chart, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category. Suppose a statistics professor collects information about the classification of her students about which of her students are freshmen, sophomores, juniors, or seniors. The data she collects is summarized in the pie chart below.

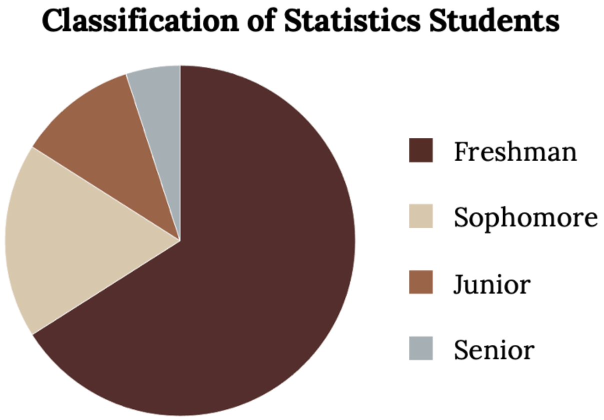


Figure 2.5: Classification of statistics students. Figure description available at the end of the section.

Bar Graphs

Bar graphs are made up of separate bars representing categories, where the length of the bar for each category is proportional to the number or percent of individuals in that category. The bars can be rectangles, or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. In the bar graph shown in Figure 2.7 below, age groups are represented on the x-axis and proportions on the y-axis

By the end of 2011, Facebook had over 146 million users in the United States. The figure below shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Figure 2.6: Facebook users

A bar graph of this data would look as follows:

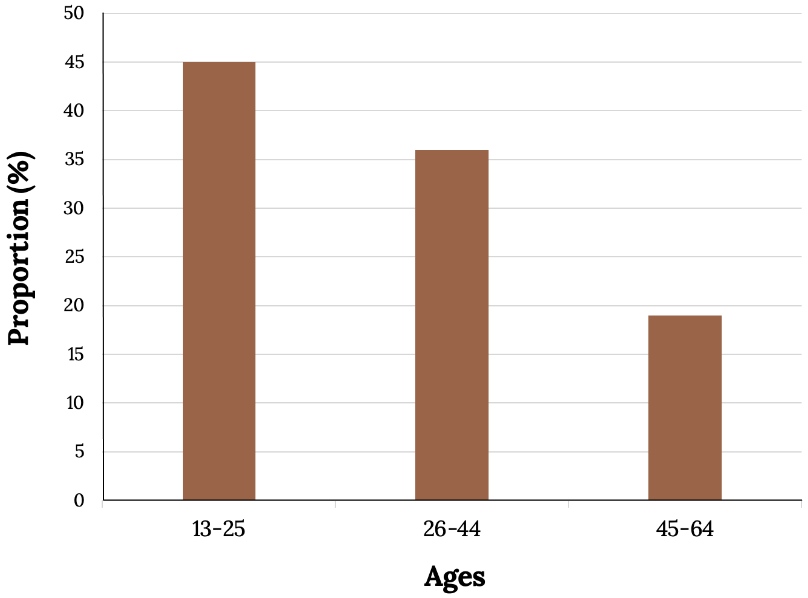


Figure 2.7: Facebook users (bar graph). [Figure description available at the end of the section](#).

Pie vs. Bar Charts

It is a good idea to consider a variety of graphs to determine which will be the most helpful in displaying our data. Our choice of the "best" graph will change depending on the data and the context. Our choice also depends on our purpose behind using the data. Look at the following plots (pie or bar), and think about which displays the comparisons better:

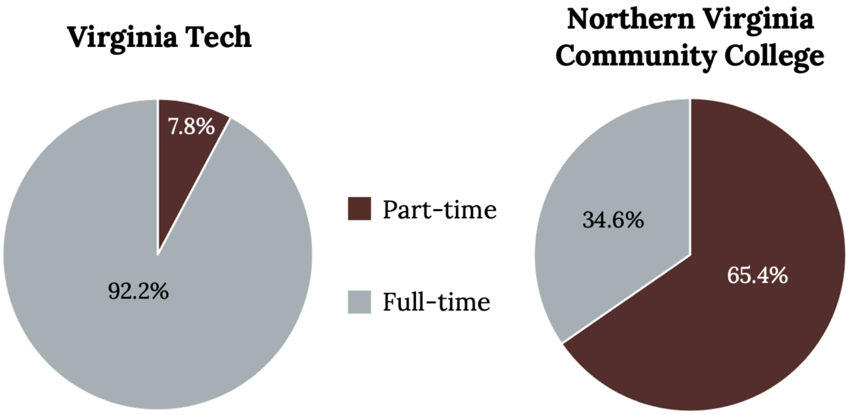


Figure 2.8: Full-time and part-time students at Virginia Tech and NVCC (pie chart). [Figure description available at the end of the section](#).

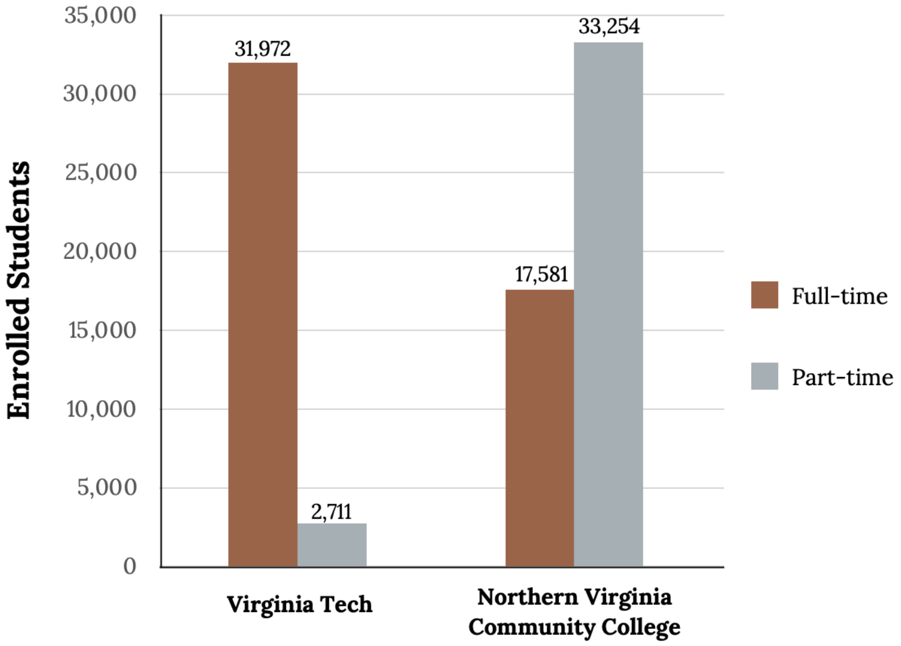


Figure 2.9: Full-time and part-time students at Virginia Tech and NVCC (bar graph). Figure description available at the end of the section.

Percentages That Add up to More (or Less) than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph shown in Figure 2.11, the percentages add to more than 100% because students can be in more than one category. Therefore, a bar graph is appropriate to compare the relative size of the categories, and a pie chart cannot be used. It also cannot be used if the percentages add up to less than 100%.

Characteristic/Category	Percent
Full-time students	40.9%
Students who intend to transfer to a four-year educational institution	48.6%
Students under age 25	61.0%
Total	150.5%

Figure 2.10: De Anza College data

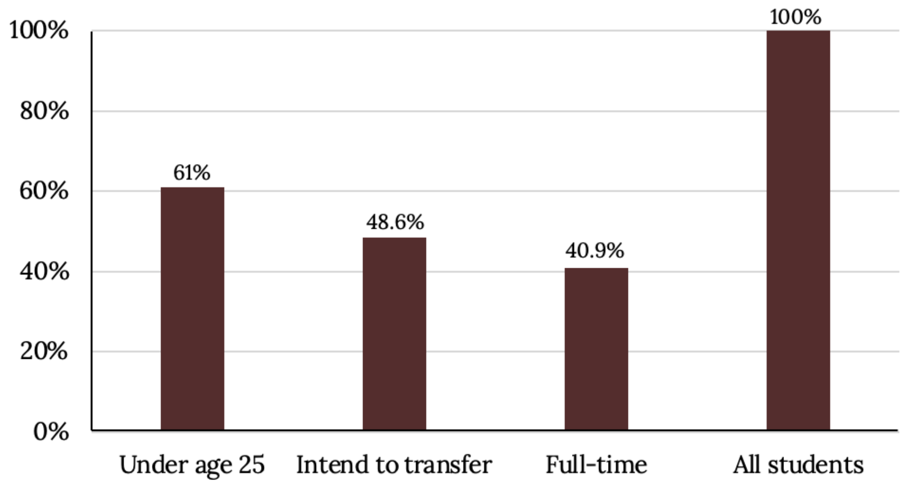


Figure 2.11: De Anza College bar graph. Figure description available at the end of the section.

Omitting Categories/Missing Data

The table displays ethnicity of students but is missing the "Other/Unknown" category for people who did not feel they fit into any of the ethnicity categories or who declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
Total	22,044 out of 24,382	90.4% out of 100%

Figure 2.12: Ethnicity of students at De Anza College

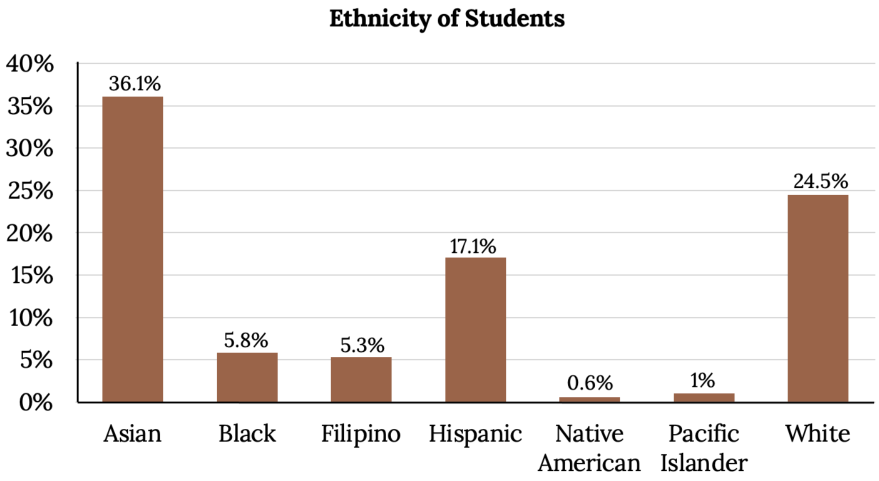


Figure 2.13: Ethnicity of students at De Anza College (bar graph). Figure description available at the end of the section.

The graph shown in Figure 2.14 is the same as the previous graph, but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories, like Native American (0.6%) or Pacific Islander (1.0%). This is important to know when we think about what the data is telling us.

Bar Graph with "Other/Unknown" Category

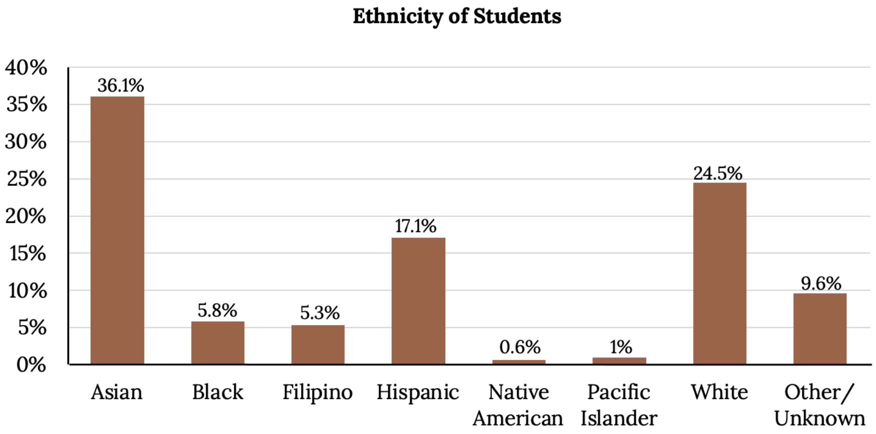


Figure 2.14: Ethnicity of students at De Anza College (bar graph with "Other/ Unknown"category). Figure description available at the end of the section.

This particular bar graph could be difficult to understand visually at first glance. A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest). This Pareto chart arranges the bars from largest to smallest and is easier to read and interpret.

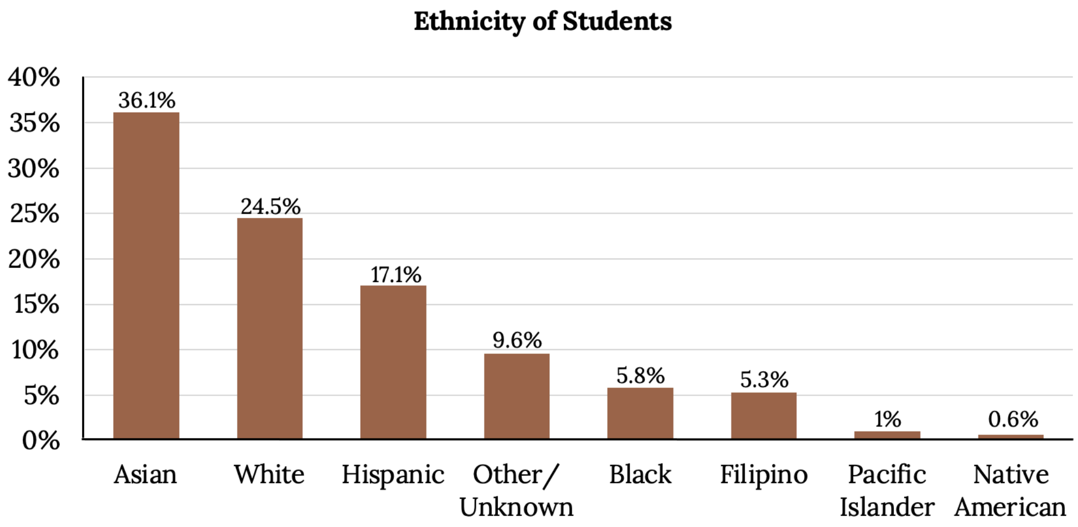


Figure 2.15: Ethnicity of students at De Anza College (bar graph with "Other/Unknown" category). *Figure description available at the end of the section.*

Pie Charts: No Missing Data

The following pie charts include the "Other/Unknown" category (since the percentages must add to 100%). The chart on the right is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in the chart on the left.

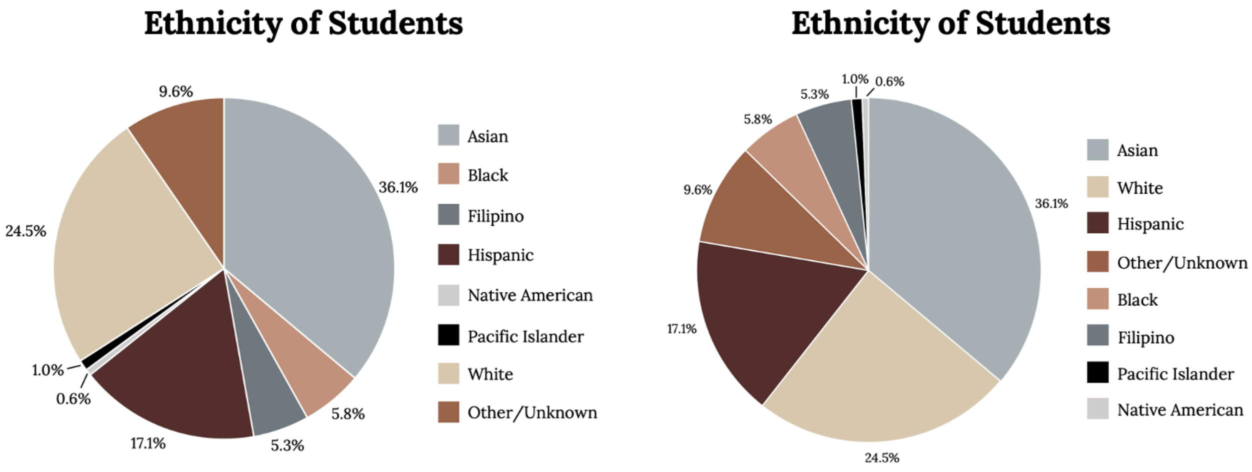


Figure 2.16: Pie charts with no missing data. *Figure description available at the end of the section.*

Example

The columns in the table below contain the race or ethnicity of students in U.S. public schools for the class of 2011, percentages of that class taking the Advanced Placement exam, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the x-axis and the Advanced Placement examinee population percentages on the y-axis.

Race/Ethnicity	AP examinee population	Overall student population
1 = Asian, Asian American, or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Figure 2.17: AP student population

Solution:

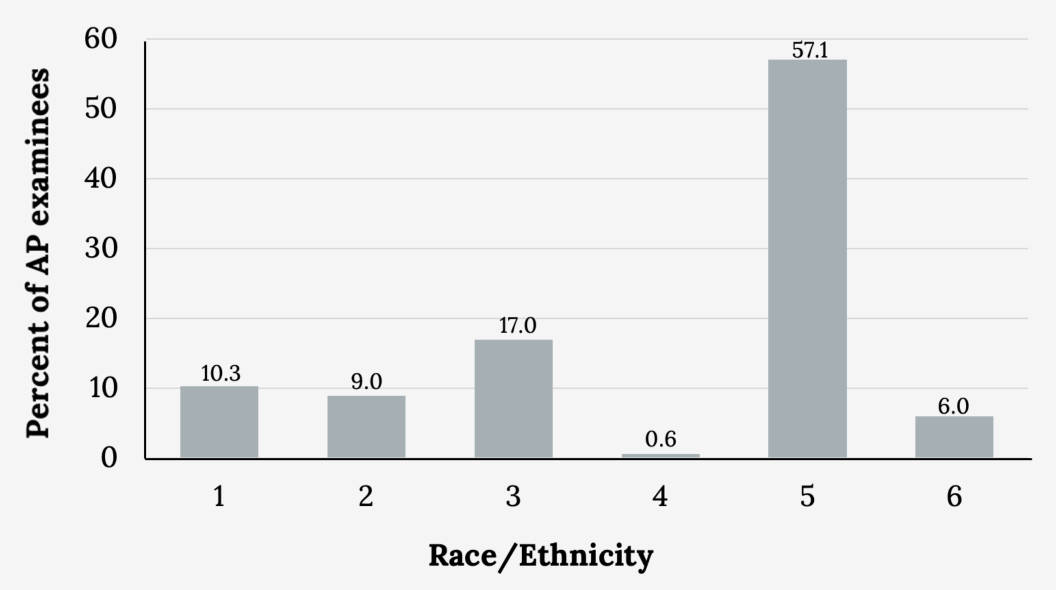


Figure 2.18: AP student population (bar graph). Figure description available at the end of the section.

Your Turn!

Park City is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Figure 2.19: Registered voter population by district

Construct a bar graph that shows the registered voter population by district.

Describing Categorical Data

After we have displayed the data visually, we then want to follow up by describing it numerically. Since categorical data does not lend itself to mathematical calculations by nature, there are not many numerical descriptors we can use to describe it. However, we can describe a categorical distribution's "typical value" with the **mode**, as well as noting its level of **variability.**

Mode

The mode of a dataset is the most frequently occurring value. There can be more than one mode in a dataset as long as those values have the same frequency and that frequency is the highest. A dataset with two modes is called bimodal, and one with three modes is called trimodal. Sets with multiple modes are called multimodal. In most cases, the mode can easily be found as the largest piece of a pie chart or largest bar in a bar chart. Modes can be observed in several previous examples from this chapter:

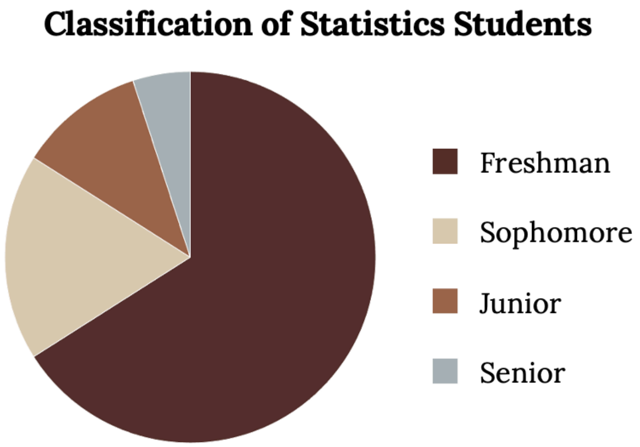


Figure 2.20: Classification of statistics students. Figure description available at the end of the section.

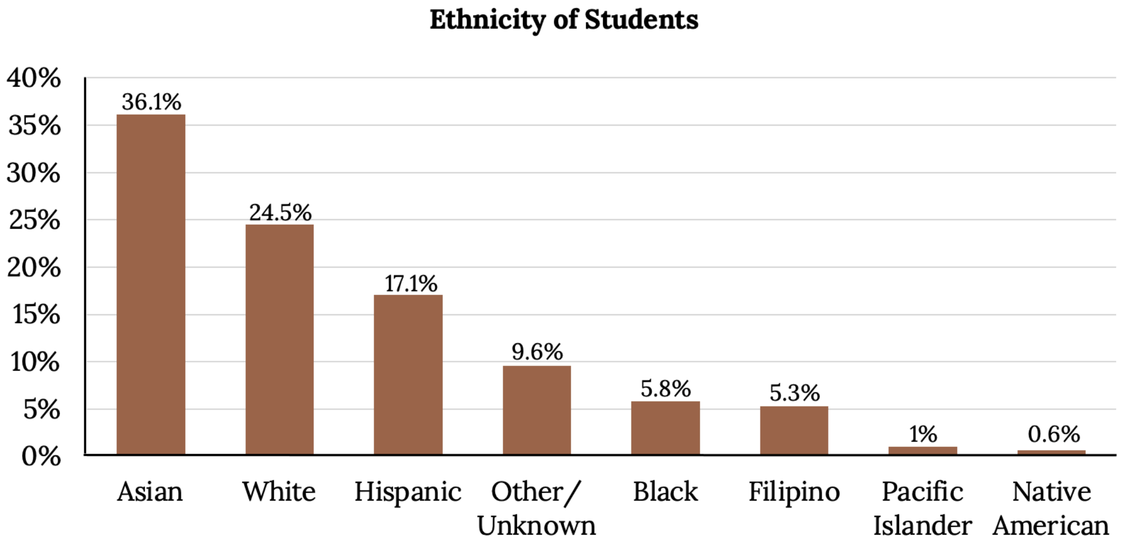


Figure 2.21: Ethnicity of students at De Anza College (bar graph with "Other/Unknown" category). Figure description available at the end of the section.

The mode of the class of statistics students shown in Figure 2.20 is obviously Freshman. If any doubt remains, a Pareto chart makes identifying the mode trivial, which is Asian in Figure 2.21.

Variability

The best way to gauge variability in categorical data is by thinking about it as *diversity*. Although we will not calculate a numerical measure here, we can note it visually. A variable that has observations spread out fairly evenly over all categories shows high variability, while a variable with observations mostly falling into one or a small number of categories displays low variability.

Example

Consider the level of variability in the two pie charts below. Which college has more variability?

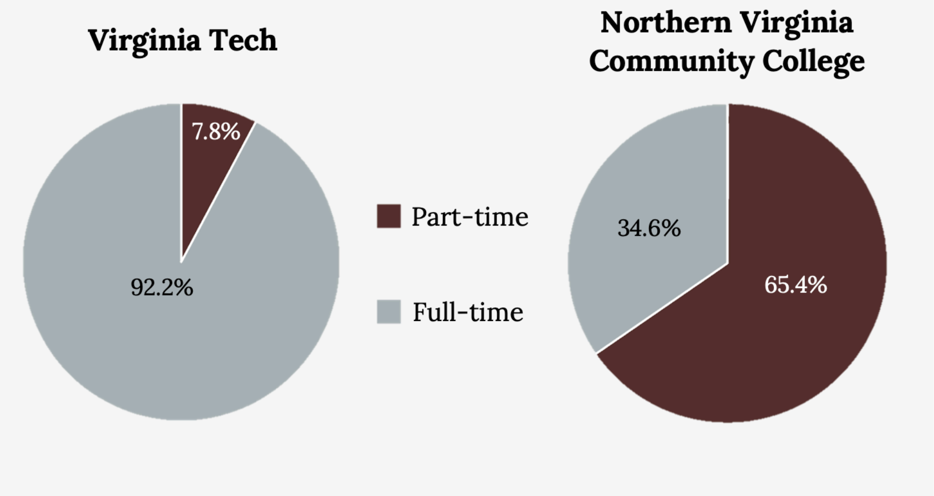


Figure 2.22: Full-time and part-time students at Virginia Tech and NVCC. Figure description available at the end of the section.

Solution
Although this variable only has two levels, Northern Virginia Community College shows more variability than Virginia Tech with regard to types of students.

Your Turn!

Let's consider the variability in the following bar charts. Which bar chart shows greater variability?

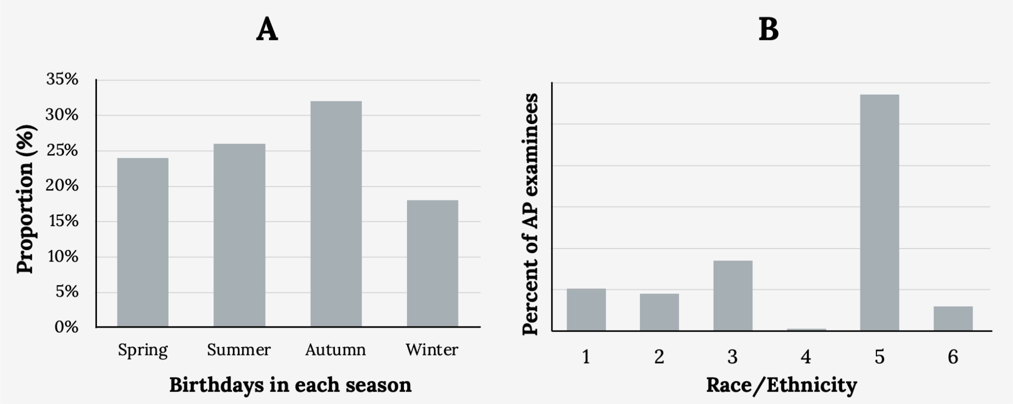


Figure 2.23: Variability comparisons. Figure description available at the end of the section.

Solution

We see much greater variability in the birthday data on the left than the ethnicity data on the right.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 2.5: Kindred Grey (2020). *Classification of statistics students.* CC BY-SA 4.0.

Figure 2.7: Kindred Grey (2020). *Facebook users (bar graph).* CC BY-SA 4.0.

Figure 2.8: Kindred Grey (2020). *Full-time and part-time students at Virginia Tech and NVCC (pie chart).* CC BY-SA 4.0.

Figure 2.9: Kindred Grey (2020). *Full-time and part-time students at Virginia Tech and NVCC (bar graph).* CC BY-SA 4.0.

Figure 2.11: Kindred Grey (2020). *De Anza College bar graph.* CC BY-SA 4.0.

Figure 2.13: Kindred Grey (2020). *Ethnicity of students at De Anza College (bar graph).* CC BY-SA 4.0.

Figure 2.14: Kindred Grey (2020). *Ethnicity of students at De Anza College (bar graph with "Other/Unknown"category).* CC BY-SA 4.0.

Figure 2.15: Kindred Grey (2020). *Ethnicity of students at De Anza College (bar graph with "Other/Unknown" category).* CC BY-SA 4.0.

Figure 2.16: Kindred Grey (2020). *Pie charts with no missing data.* CC BY-SA 4.0.

Figure 2.18: Kindred Grey (2020). *AP student population (bar graph).* CC BY-SA 4.0.

Figure 2.20: Kindred Grey (2020). *Classification of statistics students.* CC BY-SA 4.0.

Figure 2.21: Kindred Grey (2020). *Ethnicity of students at De Anza College (bar graph with "Other/Unknown" category).* CC BY-SA 4.0.

Figure 2.22: Kindred Grey (2020). *Full-time and part-time students at Virginia Tech and NVCC.* CC BY-SA 4.0.

Figure 2.23: Kindred Grey (2020). *Variability comparisons.* CC BY-SA 4.0.

Figure Descriptions

Figure 2.5: Pie chart showing the class classification of statistics students. The chart has four sections labeled Freshman, Sophomore, Junior, Senior with associated descending pie slice sizes.

Figure 2.7: Bar graph that matches the supplied data. The x-axis shows age groups (13-25, 26-44, 45-64), and the y-axis shows the proportion (%) of Facebook users ranging from zero to 50.

Figure 2.8: Two pie charts. Left pie chart labeled Virginia Tech separated into two pie slices: Part time (7.8%) and Full time (92.2%). Right pie chart labeled Northern Virginia Community College separated into two pie slices: Part time (65.4%) and Full time (34.6%).

Figure 2.9: Y axis range from zero to 35,000, X axis shows: Virginia Tech full time (31972), Virginia Tech part

time (2711), Northern Virginia Community College full time (17581), Northern Virginia Community College part time (33254).

Figure 2.11: Bar graph with Y axis ranging from 0% to 100% by 20%. X axis values: Under age 25 (61%), Intend to transfer (48.6%), full time (40.9%), all students (100%).

Figure 2.13: Bar graph with Y axis ranging from 0% to 40% by 5%. X axis values: Asian (36.1%), Black (5.8%), Filipino (5.3%), Hispanic (17.1%), Native American (0.6%), Pacific Islander (1%), White (24.5%).

Figure 2.14: Bar graph with Y axis ranging from 0% to 40% by 5%. X axis values: Asian (36.1%), Black (5.8%), Filipino (5.3%), Hispanic (17.1%), Native American (0.6%), Pacific Islander (1%), White (24.5%), Other/Unknown (9.6%).

Figure 2.15: Bar chart including the same values as Figure 2.17, however this figure is sorted from highest to lowest, left to right, beginning with Asian (36.1%) and ending with Native American (0.6%).

Figure 2.16: Two pie charts side by side both titled 'Ethnicity of Students'. Left: Asian (36.1%), Black (5.8%), Filipino (5.3%), Hispanic (17.1%), Native American (0.6%), Pacific Islander (1%), White (24.5%), Other (9.6%). Right pie chart includes the same values but arranged in descending order starting with Asian (36.1%) and ending with Native American (0.6%).

Figure 2.18: Bar graph that matches the supplied data. The x-axis shows race and ethnicity with groups one through six, and the y-axis shows the percentages of AP examinees ranging from 0.6 to 57.1.

Figure 2.20: Pie chart showing the class classification of statistics students. The chart has four sections labeled Freshman, Sophomore, Junior, Senior with associated descending pie slice sizes.

Figure 2.21: Bar chart including the same values as Figure 2.17, however this figure is sorted from highest to lowest, left to right, beginning with Asian (36.1%) and ending with Native American (0.6%).

Figure 2.22: Two pie charts. Left pie chart labeled Virginia Tech separated into two pie slices: Part time (7.8%) and Full time (92.2%). Right pie chart labeled Northern Virginia Community College separated into 2 pie slices: Part time (65.4%) and Full time (34.6%).

Figure 2.23: Two bar graphs side by side. Left: titled 'A' with 'Proportion (%) on the y axis and 'Birthdays in each season' on the x axis. X axis includes: Spring (24%), Summer (26%), Autumn (31%), Winter (18%). Right: titled 'B' with 'Percent of AP examinees' on the y axis and 'Race/Ethnicity' on the x axis. X axis values: 1 (10.3), 2 (9.0), 3 (17.0), 4 (0.6), 5 (57.1), 6 (6.0).

2.3 Displaying Quantitative Distributions

Descriptive Statistics for Quantitative Data

Descriptive options for **quantitative data** are much more robust than for categorical. As previously mentioned, descriptive statistics can be expressed both visually and numerically (usually in that order).

This section will expand on graphical methods, while the next few sections will focus on numerical summaries of quantitative data.

Graphical Methods for Quantitative Data

The first thing we may do, especially for quantitative data, is to examine it in a frequency table. We have many more graphical options beyond that for quantitative data, including:

- Stem-and-leaf plots
- Dot plots
- Line graphs
- Histograms
- Frequency polygons
- Time series plots

Each of these methods has its own distinct advantages and disadvantages.

Stem-and-Leaf Plots

One simple graph, the stem-and-leaf graph or stemplot, comes from the field of exploratory data analysis. This graph is a good choice when datasets are small. To create the plot, divide each observation of data into a "stem" and a "leaf." The stem is the first part of the number, while the leaf consists of a final significant digit. For example, you could divide the number 23 into a stem of 2 and a leaf of 3. The number 432 could have a stem of 43 and leaf of 2. The decimal 9.3 could have a stem of 9 and leaf of 3. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stems.

Example

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest): 33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100.

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

Figure 2.24: Exam 1 scores

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores, or approximately 26% $\left(\frac{8}{31}\right)$, were in the 90s or 100, a fairly high number of As.

The stemplot is a quick way to organize things and gives a good picture of the data. You can quickly and easily find basic summary statistics such as the maximum, minimum, range, and some measures we will explore in the future, such as the median and quartiles. Stemplots can be good for seeing individual data points and mainly handle discrete or rounded continuous data.

Comparisons with Stem-and-Leaf Plots

Back-to-back or side-by-side stem-and-leaf plots allow for the comparison of two datasets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem, with one set on the left and one on the right.

Your Turn!

The following two tables show the ages of U.S. presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

President	Age	President	Age	President	Age	President	Age
Washington	57	Fillmore	50	McKinley	54	Nixon	56
J. Adams	61	Pierce	48	T. Roosevelt	42	Ford	61
Jefferson	57	Buchanan	56	Taft	51	Cater	52
Madison	57	Lincoln	52	Wilson	56	Reagan	69
Monroe	58	A. Johnson	56	Harding	55	G. H. W. Bush	64
J. Q. Adams	57	Grant	46	Coolidge	51	Clinton	47
Jackson	61	Hayes	54	Hoover	54	G. W. Bush	54
Van Buren	55	Garfield	49	F. Roosevelt	51	Obama	47
W. H. Harrison	68	Arthur	51	Truman	60	Trump	70
Tyler	51	Cleveland	47	Eisenhower	62	Biden	78
Polk	49	B. Harrison	55	Kennedy	43		
Taylor	64	Cleveland	55	L. Johnson	55		

Figure 2.25: Presidential ages at inauguration

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58	G. H. W. Bush	94
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Figure 2.26: Presidential ages at death

Solution

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 8 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
8 0	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3 4

Line Graphs

Another type of graph that is useful for showing trends in specific data values (i.e., **discrete** data) is a line graph. In the particular line graph shown below, the x-axis (horizontal axis) consists of data values and the y-axis (vertical axis) consists of frequency points. The frequency points are connected using line segments.

NOTE: Line graphs can also be used with some **ordinal categorical data**.

Example

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to complete chores. The results are shown in the table and chart below.

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

Figure 2.27: Chore reminder data

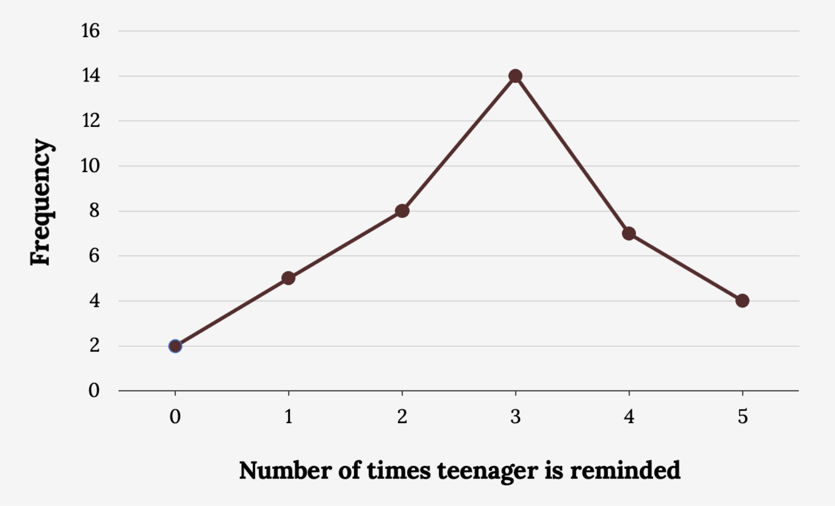


Figure 2.28: Chore reminder (line graph). Figure description available at the end of the section.

Dot Plots

A dot plot consists of a number line and dots (or points) positioned above the number line.

Dot plots are very similar in functionality to stem-and-leaf plots but look a little bit cleaner. They can reveal an overall pattern and any outliers or extreme values. An outlier is an observation of data that does not fit the rest of the data. When graphed, an outlier will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500), while others may indicate that something unusual is happening. It takes some background information to fully explain outliers; we will cover them in more detail later.

Example

Consider the following data dealing with the hours of sleep students get per night: 5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9

The dot plot for this data would be as follows:

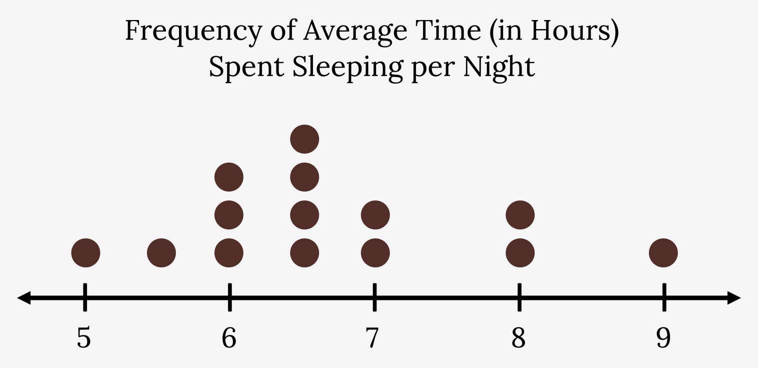


Figure 2.29: Student sleep hours. Figure description available at the end of the section.

Histograms

For most of the work in this book, histograms will display the data. One advantage of a histogram is that it can readily display large continuous datasets. A rule of thumb is to use a histogram when the dataset consists of 100 or more values.

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either “frequency” or “relative frequency” (or “percent frequency” or “probability”). The graph will have the same shape regardless of label. The histogram can give you a really good look at the overall shape of the data, the center, and the spread. However, you do lose individual data points.

A histogram is essentially a two-dimensional frequency table. To construct a histogram, you must first decide the size and number of bars, intervals, or classes, similarly to how you would with a frequency table.

Example

The following data are the heights (in inches to the nearest half-inch) of 100 male semiprofessional soccer players. The heights are continuous data, since height is measured.

60, 60.5, 61, 61, 61.5, 63.5, 63.5, 63.5, 64, 64, 64, 64, 64, 64, 64, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5, 68, 68, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69.5, 69.5, 69.5, 69.5, 69.5, 70, 70, 70, 70, 70, 70, 70.5, 70.5, 70.5, 71, 71, 71, 72, 72, 72, 72.5, 72.5, 73, 73.5, 74

The smallest data value is 60. Since none of the data has more than one decimal, we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

This results in 60 – 0.05 = 59.95, which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$\frac{74.05-59.95}{8}=1.76$$

NOTE:

We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. In determining the number of bars or class intervals, some follow the guideline to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

Some values in datasets might fall on boundaries for different intervals. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the x-axis and relative frequency on the y-axis.

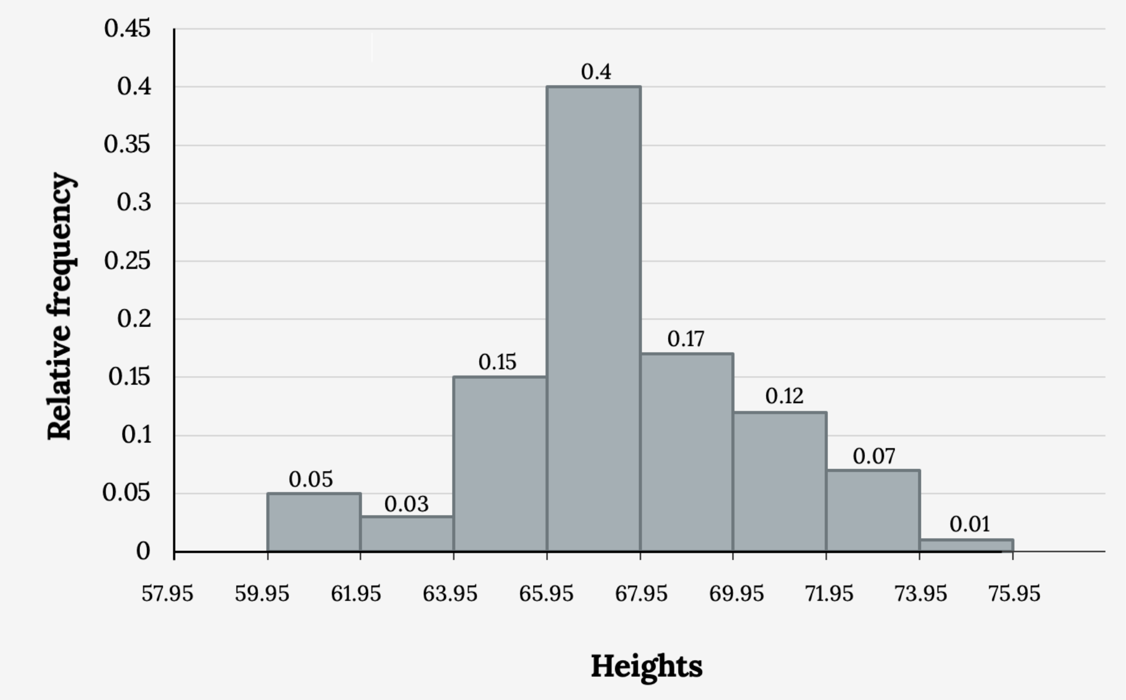


Figure 2.30: Soccer player heights. *Figure description available at the end of the section.*

Frequency Polygons

Frequency polygons are analogous to line graphs but instead utilize binning techniques to make continuous data visually easy to interpret. They are essentially combinations of histograms and line graphs.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x-axis and y-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Frequency polygons are sometimes more useful than histograms for comparing continuous distributions. This is achieved by overlaying the frequency polygons drawn for different datasets.

Example

A frequency polygon was constructed from the frequency table below.

Lower bound	Upper bound	Frequency	Cumulative frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
89.5	89.5	40	85
99.5	99.5	15	100

Figure 2.31: Frequency distribution for calculus final test scores

The first label on the x-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the x-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals, with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the x-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

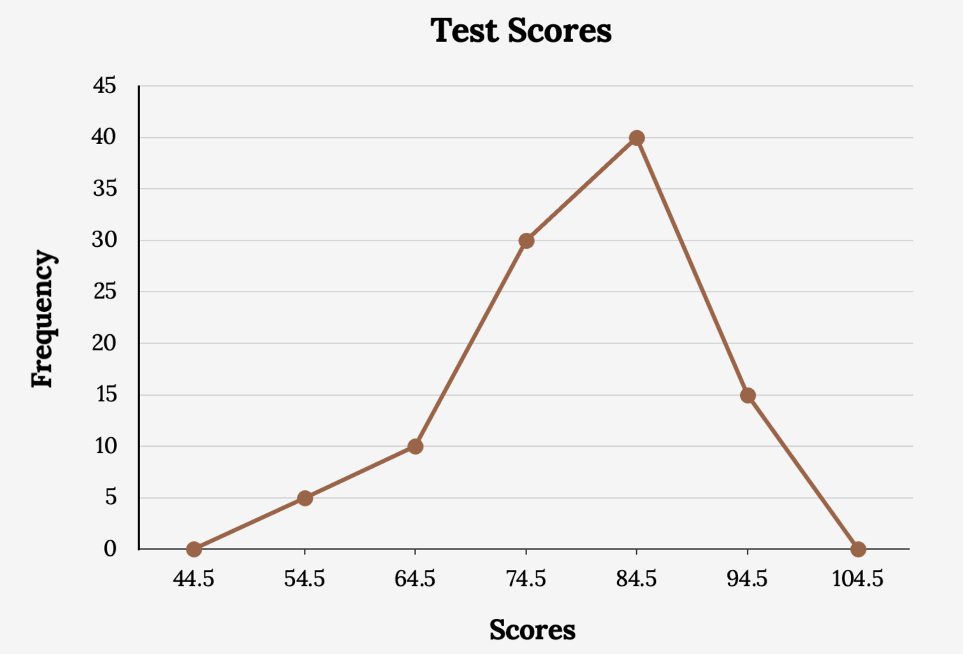


Figure 2.32: Calculus final test scores (frequency polygon). *Figure description available at the end of the section.*

Time Series Plots

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon, we take note of the temperature in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reached a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is time. Since each date is paired with the temperature reading for the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses would be a time series graph.

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

To construct a time series graph, we must look at both pieces of our paired dataset. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot dates or other time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example

The following data shows the Annual Consumer Price Index each month for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	January	Feburary	March	April	May	June	July
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.655	227.663	229.392	230.085	229.815	229.478	229.104
2013	230.280	232.166	232.773	232.531	232.945	233.504	233.596
2014	233.916	234.781	236.293	237.072	237.900	238.343	238.250
2015	233.707	234.722	236.119	236.599	237.805	238.638	238.654
2016	236.916	237.111	238.132	239.261	240.236	241.038	240.647
2017	242.839	243.603	243.801	244.524	244.733	244.955	244.786
2018	247.867	248.991	249.554	250.546	251.588	251.989	252.006
2019	251.712	252.776	254.202	255.548	256.092	256.143	256.571

Year	August	September	October	November	December	Annual
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594
2013	233.877	234.149	233.546	233.069	233.049	232.957
2014	237.852	238.031	237.433	236.151	234.812	236.736
2015	238.316	237.945	237.838	237.336	236.525	237.017
2016	240.853	241.428	241.729	241.353	241.432	240.007
2017	245.519	246.819	246.663	246.669	246.524	245.120
2018	252.146	252.439	252.885	252.038	251.233	251.107
2019	256.558	256.759	257.346	257.208	256.974	255.657

Figure 2.33: CPI data

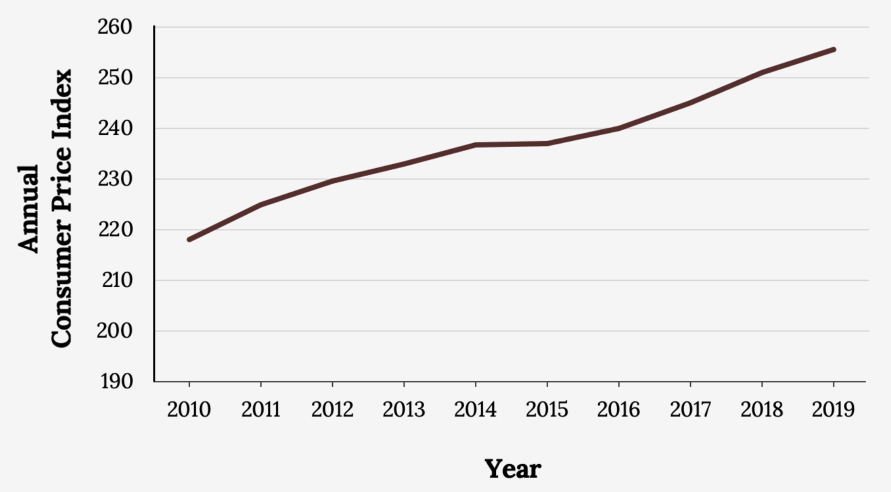


Figure 2.34: CPI time series plot. *[Figure description available at the end of the section](#)*.

Additional Resources

[Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.](#)

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 2.28: Kindred Grey (2020). *Chore reminder (line graph)*. CC BY-SA 4.0.

Figure 2.29: Kindred Grey (2020). *Student sleep hours*. CC BY-SA 4.0.

Figure 2.30: Kindred Grey (2020). *Soccer player heights*. CC BY-SA 4.0.

Figure 2.32: Kindred Grey (2020). *Calculus final test scores (frequency polygon)*. CC BY-SA 4.0.

Figure 2.33: Data retrieved from https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008

Figure 2.34: Kindred Grey (2020). *CPI time series plot*. CC BY-SA 4.0.

Figure Descriptions

Figure 2.28: Line graph showing the number of times a teenager needs to be reminded to do chores on the x-axis (range one through six by one) and frequency on the y-axis (range zero through 16 by two).

Figure 2.29: Dot plot showing 'frequency of average time (in hours) spent sleeping per night'. The number line is marked in intervals of one from five to nine. Dots above the line show one person reporting five hours, one with 5.5, three with 6, four with 6.5, two with 7, two with 8, and one with 9 hours.

Figure 2.30: Histogram consists of eight bars with the y-axis in increments of 0.05 from 0-0.45 measuring relative frequency and the x-axis in intervals of two from 57.95-75.95 measuring heights. The highest is 65.95-67.95 (0.4 relative frequency).

Figure 2.32: X axis measures scores and the y axis measures frequency. Highest frequency is 40 when the score is 84.5.

Figure 2.34: Times series graph that matches the supplied data. The x-axis shows years from 2010 to 2019, and the y-axis shows the annual CPI. Constant positive trend.

2.4 Describing Quantitative Distributions

Consider the following exercise.

Your classmates write down the average time (in hours, to the nearest half-hour) they sleep per night and then create a simple dot plot of the data:

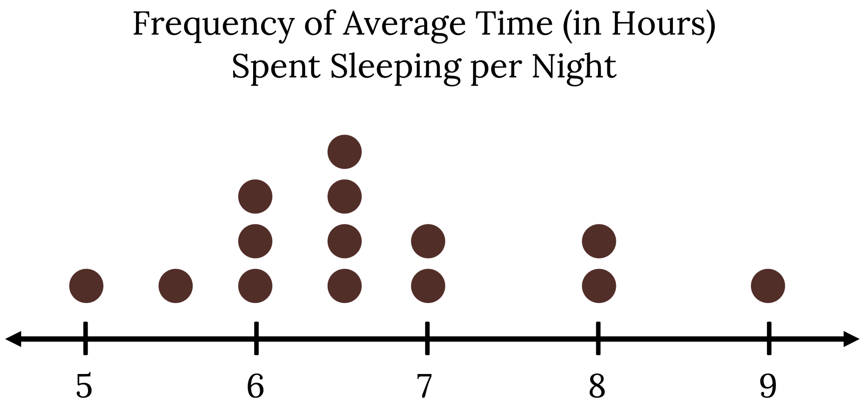


Figure 2.35: Student sleep hours. [Figure description available at the end of the section](#).

How would you interpret or explain this distribution? Where do your data appear to cluster? How might you interpret the clustering? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

The questions above ask you to analyze and interpret your data. It isn't enough to just make graphs, we must be able to interpret the information with a critical eye.

Key Aspects of Quantitative Data

When describing a quantitative distribution we want to note at least four: the shape of the distribution, the presence of outliers, the center, and the spread. A helpful acronym for remembering this is **SOCS**:

- **Shape**
- **Outliers**
- **Center**
- **Spread**

Shape is the main characteristic we can determine by looking at a graph. We are often able to identify potential outliers visually as well. Center and spread can be roughly gauged visually, but there are also numerical calculations for them, which will be discussed in the following sections.

Shape

Shape is the first thing we should note since it will often dictate how to proceed with the rest of our analysis. We have already seen that most of our graphical methods can give us an idea of the shape of a distribution, but the best choice in most situations is a properly formatted histogram.

Symmetry vs. Skewness

Most of us are familiar with datasets that show roughly equal tails trailing off equally in both directions, which would be described as symmetric.

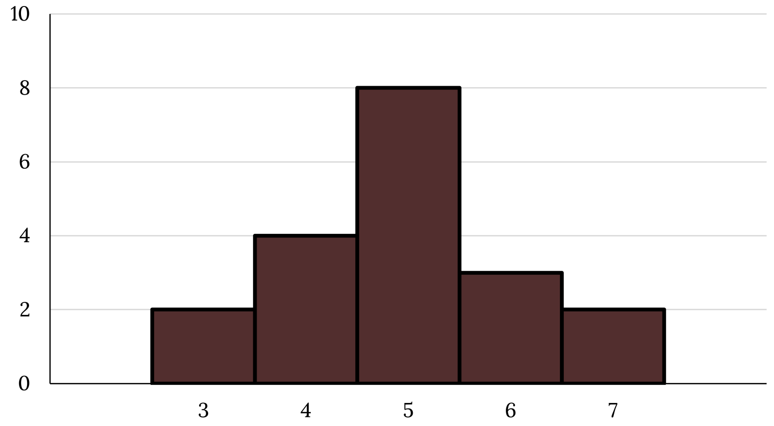


Figure 2.36: Symmetric data. [*Figure description available at the end of the section*](#).

Consider the following alternative:

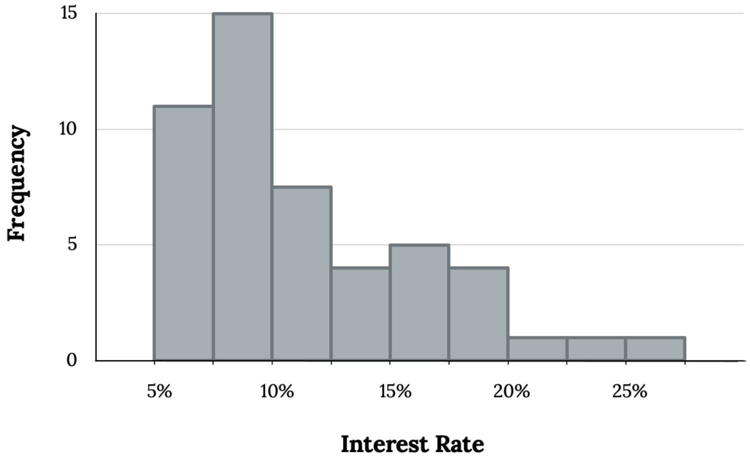


Figure 2.37: Skewed data. *[Figure description available at the end of the section](#).*

The figure above suggests that most loans have rates under 15%, while only a handful of loans have rates above 20%. When data trails off to the right in this way and has a longer right tail, the shape is said to be right-skewed. Datasets with the reverse characteristic—a long, thinner tail to the left—are said to be left-skewed. We also say that such a distribution has a long left tail.

Modality

In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify the **modality** of a distribution. A mode is represented by a prominent peak in the distribution. There is only one prominent peak in the histogram of loan amounts. The definition of mode sometimes taught in math classes is the value with the most occurrences in the dataset. However, for many real-world datasets, it is common to have no observations with the same value in a dataset, making this definition impractical in data analysis. The figure below shows histograms that have one, two, or three prominent peaks.

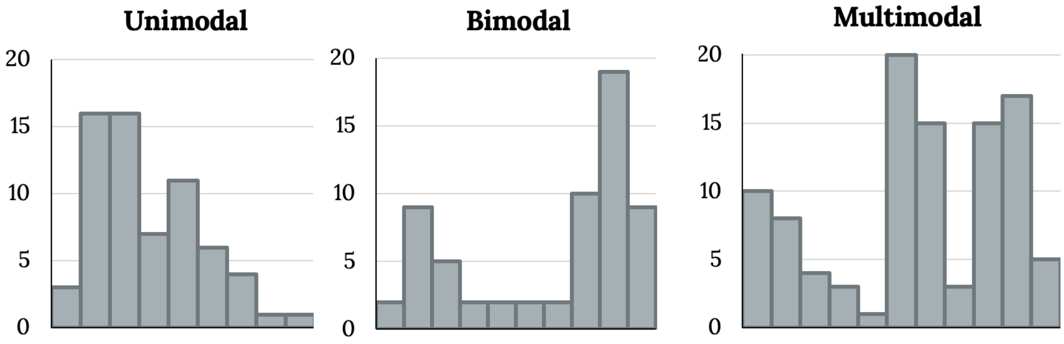


Figure 2.38: Unimodal, bimodal, and multimodal distributions. *[Figure description available at the end of the section](#).*

Such distributions are called unimodal, bimodal, and multimodal, respectively. Any distribution with more than two prominent peaks is called multimodal. Notice that there was one prominent peak in the unimodal distribution, with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations.

Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why "prominent" is not rigorously defined in this book. The most important part of this examination is to better understand your data.

Outliers

Sometimes one or more data points stick out visually. These extreme values could potentially be **outliers**. Sometimes they may be obvious to us, as in the following histogram:

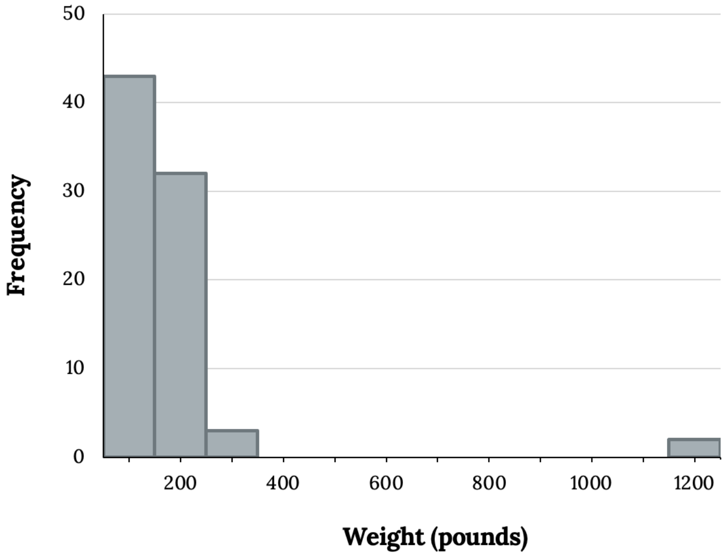


Figure 2.39: Outlier. *Figure description available at the end of the section.*

On the other hand, outliers may not be as obvious and might only show up upon careful examination of a dot plot or through other methods. Examining data for outliers serves many useful purposes, including:

- Identifying skewness in the distribution
- Identifying possible data collection or data entry errors
- Providing insight into interesting properties of the data

Subsequent sections will discuss numerical methods to "officially" identify outliers and how to deal with them.

Center

We also want to make sure to describe a quantitative distribution's most typical value, known as its central tendency. We can simply estimate this visually, but future sections will focus on more robust and appropriate measures we can calculate.

Spread

A rough measure of spread we can usually determine visually is the range (range = maximum − minimum). Again, we will encounter more robust and appropriate measures we can calculate in future sections.

Example

Use the following graph to answer the questions below.

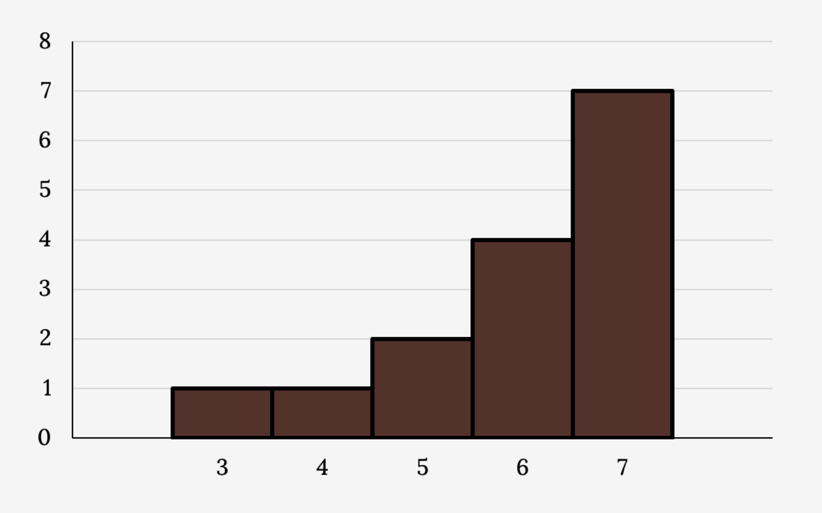


Figure 2.40: Distribution 1. Figure description available at the end of the section.

Describe the shape of this distribution.

Solution
The distribution is left-skewed.

Describe the modality of the distribution.

Solution
The distribution appears to be unimodal with a mode of 7.

Do you see any apparent outliers?

Solution
3 may be a potential outlier.

What does the center appear to be?

Solution
The center appears to be roughly 6.

Provide a rough estimate of the spread.

Solution
We see a range of 7 – 3 = 4 for a rough measure of spread.

Your Turn!

Describe the shape of this distribution visually:

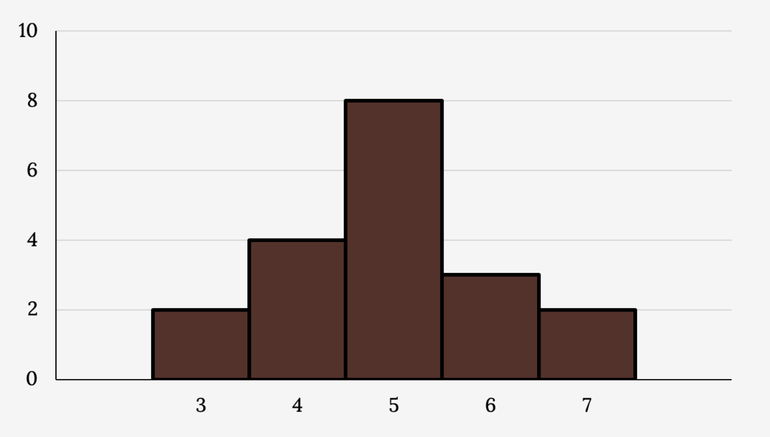


Figure 2.41: Distribution 2. Figure description available at the end of the section.

- Is the data symmetric or skewed? If you see skewness, what is its direction?
- Describe the modality of the distribution.

- Do you see any apparent outliers?
- What does the center appear to be?
- Provide a rough estimate of the spread.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 2.35: Kindred Grey (2020). *Student sleep hours.* CC BY-SA 4.0.

Figure 2.36: Kindred Grey (2020). *Symmetric data.* CC BY-SA 4.0.

Figure 2.37: Kindred Grey (2020). *Skewed data.* CC BY-SA 4.0.

Figure 2.38: Kindred Grey (2020). *Unimodal, bimodal, and multimodal distributions.* CC BY-SA 4.0.

Figure 2.39: Kindred Grey (2020). *Outlier.* CC BY-SA 4.0.

Figure 2.40: Kindred Grey (2020). *Distribution 1.* CC BY-SA 4.0.

Figure 2.41: Kindred Grey (2020). *Distribution 2.* CC BY-SA 4.0.

Figure Descriptions

Figure 2.35: Dot plot showing 'frequency of average time (in hours) spent sleeping per night'. The number line is marked in intervals of one from five to nine. Dots above the line show one person reporting 5 hours, one with 5.5, three with 6, four with 6.5, two with 7, two with 8, and one with 9 hours.

Figure 2.36: Bell curve shaped histogram. Tallest bar is in the middle and tapers off on both sides.

Figure 2.37: Bar graph with frequency on the y axis ranging from zero to 15 by five, and Interest rate on the x axis ranging from 5% to 25% by five. Bars include: 5-7.5% (11), 7.5-10% (15), 10-12.5% (7), 12.5-15% (4), 15-17.5% (5), 17.5-20% (5), 20-22.5% (1), 22.5-25% (1), 25-27.5% (1).

Figure 2.38: Three side-by-side bar graphs with y axis ranging from zero to 20 by five. The first graph has a peak around four, the second graph has a peak at three and 17, the third graph has a peak at one, 12, and 18.

Figure 2.39: Bar graph with x axis measuring weight (in pounds) ranging from zero to 1200 by 200. The y axis measures frequency ranging from zero to 40 by 10. There is a peak when weight = 200, frequency = 40. the only other data is when weight = 1200, frequency = 2.

Figure 2.40: Histogram which consists of five adjacent bars over an x-axis split into intervals of one from three to seven. The bar heights from left to right are: 1, 1, 2, 4, 7.

Figure 2.41: Histogram which consists of five adjacent bars with the x-axis split into intervals of one from three to seven. The bar heights peak in the middle and taper down to the right and left.

2.5 Measures of Location and Outliers

Let's keep working through the acronym SOCS for describing key aspects of our data, this time focusing on outliers.

- Shape
- **Outliers**
- Center
- Spread

Measures of location are used to quantify where an observation stands in relation to the rest of the distribution. They also provide the building blocks to formally identify outliers. Common measures of location are quartiles and percentiles. Quartiles divide ordered data into quarters, while percentiles divide ordered data into hundredths.

Percentiles

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively, such as when they use SAT results to determine a minimum test score that will be an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

To score in the 90th percentile of an exam does not necessarily mean that you received 90% on a test. It means that 90% of test scores are the same or less than your score and that 10% of test scores are the same or greater than your test score.

Percentiles are mostly used with very large populations. Therefore, it would be acceptable to say that 90% of the test scores are less (and not the same or less) than your score, because removing one particular data value is not significant.

There are two inverse ways you may work with percentiles: finding the *k*th percentile of a distribution or finding the percentile of a given observation.

Finding the *k*th Percentile of a Distribution

Sometimes you may want to find the "*k*th" percentile of a distribution. For instance, what would a student have to score on the SAT to be in the 90th percentile?

If you were to do a little research, you would find several processes for calculating the kth percentile. Here is one of them.

k = the kth percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$.
- If i is an integer, then the kth percentile is the data value in the ith position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered dataset. This is easier to understand in an example.

NOTE: You can calculate percentiles using calculators and computers. There are a variety of online calculators.

Example

Twenty-nine ages of winners of the Academy Award for Best Actor are listed below *in order from smallest to largest.*

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

Find the 70th percentile.

Solution
k = 70, i = the index, and n = 29
$i = \frac{k}{100}(n+1) = \frac{70}{100}(29+1) = 21$. Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

Find the 83rd percentile.

Solution
k = 83rd percentile, i = the index, and n = 29
$i = \frac{k}{100}(n+1) = \frac{83}{100}(29+1) = 24.9$, which is not an integer. Round this number down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

Your Turn!

Twenty-nine ages of winners of the Academy Award for Best Actor are listed below *in order from smallest to largest.*

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

Calculate the 20th percentile and the 55th percentile.

Finding the Percentile of a Value in a Dataset

The process for finding the corresponding percentile of a given observation is as follows:

x = the number of data values counting from the bottom of the data list up to (but not including) the data value for which you want to find the percentile

y = the number of data values equal to the data value (repeated values) for which you want to find the percentile

n = the total number of data

- Order the data from smallest to largest.
- Calculate $\frac{x+0.5y}{n}$(100). Then round to the nearest integer.

Example

Twenty-nine ages of winners of the Academy Award for Best Actor are listed below *in order from smallest to largest.*

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

Find the percentile for 58.

Solution
Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

x = 18 and y = 1.

$\frac{x+0.5y}{n}$(100) = $\frac{18+0.5(1)}{29}$(100) = 63.8. 58 is the 64th percentile.

Find the percentile for 25.

Solution
x = 3 and y = 1.
$\frac{x+0.5y}{n}$(100) = $\frac{3+0.5(1)}{29}$(100) = 12.07. 25 is the 12th percentile.

Your Turn!

Twenty-nine ages of winners of the Academy Award for Best Actor are listed below *in order from smallest to largest.*

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

Find the percentiles for 47 and 31.

Quartiles

Quartiles also deal with an ordered dataset and are really just special percentiles. The first quartile, Q_1, is the same as the 25th percentile. The second quartile, Q_2, is the same as the 50th percentile and is also called the median. The third quartile, Q_3, is the same as the 75th percentile.

The Median

The **median** is a number that measures the “halfway point” of the data. You can think of the median as the “middle value,” but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data:

1, 11.5, 6, 7.2, 4, 8, 9, 10, 6.8, 8.3, 2, 2, 10, 1

Ordered from smallest to largest:

1, 1, 2, 2, 4, 6, **6.8, 7.2**, 8, 8.3, 9, 10, 10, 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, we must interpolate, or split the difference. In this case we simply add both values together and divide by two.

$$\frac{6.8+7.2}{2} = 7$$

The median is 7. Half of the values are smaller than 7, and half of the values are larger than 7.

Depending on the context, the median could be both a measure of location and/or center. Future sections will further discuss the median and using it as a measure of center.

Finding Quartiles

Quartiles can be found by treating them either as a percentile or in a similar fashion to the median. They may or may not be part of the data. To find the quartiles, first find the median, or second quartile. The first quartile, Q$_1$, can be treated as the middle value (or median) of the lower half of the data. The third quartile, Q$_3$, can be treated as the middle value (or median) of the upper half of the data. To get the idea, consider the same dataset as above, where the median is 7.

1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

The median divides the dataset into two halves, with 7 observations in each half (pictured in different colors below). The first quartile is 2 and the third quartile is 9.

1, 1, 2, **2**, 4, 6, 6.8, 7.2, 8, 8.3, **9**, 10, 10, 11.5

In this example there are an even number of observations and we had to interpolate the median. This divided our dataset into two halves, with an odd amount of numbers in each half, meaning the quartiles were part of the data set. In other cases, if the median divides your dataset into two even halves you must interpolate the quartiles. Rather than splitting the difference, it is often more appropriate to treat Q$_1$ and Q$_3$ as the 25th and 75th percentiles, respectively.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is “good” or “bad.” The interpretation of whether a certain percentile is “good” or “bad” depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered “good”; in other contexts, a high percentile might be considered “good.” In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data but also when calculating probabilities, like those you’ll find in later chapters of this text.

NOTE:

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- Information about the context of the situation being considered
- The data value (value of the variable) that represents the percentile
- The percent of individuals or items with data values below the percentile
- The percent of individuals or items with data values above the percentile

Example

On a timed math test, the first quartile for time taken to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as it is desirable to finish a timed exam more quickly (especially since you might not finish if you take too long).

Your Turn!

For the 100-meter dash, the third quartile for time taken to finish the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Five-Number Summary

The five-number summary is a simple, easy way to quickly summarize a dataset. It is also the first step to identifying any outliers. It consists of:

1. Minimum
2. Q_1
3. Median
4. Q_3
5. Maximum

Interquartile Range

The interquartile range (IQR) is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1$$

The IQR is also helpful to determine potential **outliers** and can be used as a measure of spread.

Example

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are as follows.

0, 40, 60, 30, 60, 10, 45, 30, 300, 90, 30, 120, 60, 0, 20

The five-number summary for this dataset would look like:

- Min = 0
- Q_1 = 20
- Med = 40
- Q_3 = 60
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the IQR is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

Fence Rule

Although points may often look like outliers on a graph, we establish the **upper fence (UF)** and **lower fence (LF)** to numerically decide if a value is an outlier. The lower fence is 1.5 times the IQR subtracted from the first quartile ($LF = Q_1 - 1.5*IQR$), while the upper fence is 1.5 times the IQR added to the third quartile ($UF = Q_3 + 1.5*IQR$). If a value falls outside of these fences—i.e., less than the lower fence or greater than the upper fence—we will flag it as an outlier.

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality, or they may be a key to understanding the data. Potential outliers always require further investigation.

Example

[Continued from Sharpe Middle School example above]

The principal needs to be careful. The value 300 appears to be a potential outlier.

$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$

The value 300 is greater than 120, so it is a potential outlier. If we delete it and generate the five-number summary, we get the following values:

- Min = 0
- $Q_1 = 20$
- $Q_3 = 60$
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample, and the principal should survey more students to be sure of his survey results.

Box Plots

Box plots (also called box-and-whisker plots or box-whisker plots) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from the values of the five-number summary, which we use to compare how close they are to other data values.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values serve as the endpoints of the axis. The first quartile marks one end of the box, and the third quartile marks the other end of the box. Approximately the middle 50% of the data values fall inside the box. The "whiskers" extend from the ends of the box to the smallest and largest data values. The second quartile, or median, can be between the first and third quartiles, or it can be one, the other, or both. The box plot gives a good, quick picture of the data.

NOTE:

You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values because they have been identified as outliers according to the fence rules.

Example

Consider, again, this dataset.

1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

The first quartile is 2, the median is 7, and the third quartile is 9. The smallest value is 1, and the largest value is 11.5. The following image shows the constructed box plot.

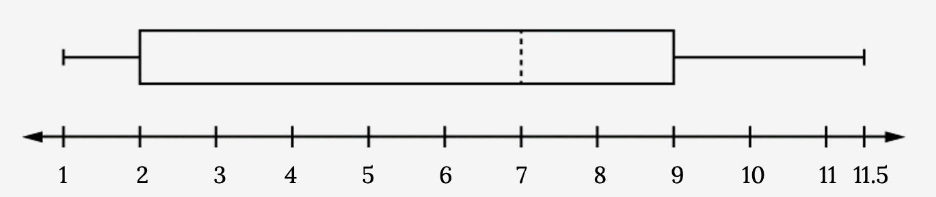


Figure 2.42: Box plot. Figure description available at the end of the section.

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

NOTE: It is important to start a box plot with a *scaled* number line. Otherwise the box plot may not be useful.

Example

The following data are the heights of 40 students in a statistics class.

59, 60, 61, 62, 62, 63, 63, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 67, 67, 68, 68, 69, 70, 70, 70, 70, 70, 71, 71, 72, 72, 73, 74, 74, 75, 77

Construct a box plot with the following properties.

- Minimum value = 59
- Maximum value = 77
- First quartile (Q1) = 64.5
- Second quartile (Q2 or median) = 66
- Third quartile (Q3) = 70

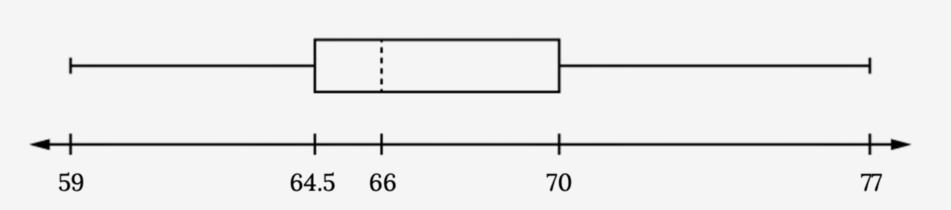


Figure 2.43: Student heights (box plot). Figure description available at the end of the section.

a. Each quarter has approximately 25% of the data.
b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread, and the fourth quarter has the largest spread.
c. Range: 77 – 59 = 18 (maximum value – minimum value)
d. Interquartile range: 70 – 64.5 = 5.5 (Q3 – Q1)
e. The interval 59–65 has more than 25% of the data, so it has more data in it than the interval 66–70, which has 25% of the data.
f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Your Turn!

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136, 140, 178, 190, 205, 215, 217, 218, 232, 234, 240, 255, 270, 275, 290, 301, 303, 315, 317, 318, 326, 333, 343, 349, 360, 369, 377, 388, 391, 392, 398, 400, 402, 405, 408, 422, 429, 450, 475, 512

For some sets of data, some values (i.e., the largest value, smallest value, first quartile, median, or third quartile) may be the same. For instance, you might have a dataset in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the largest value was 7, the box plot would look like:

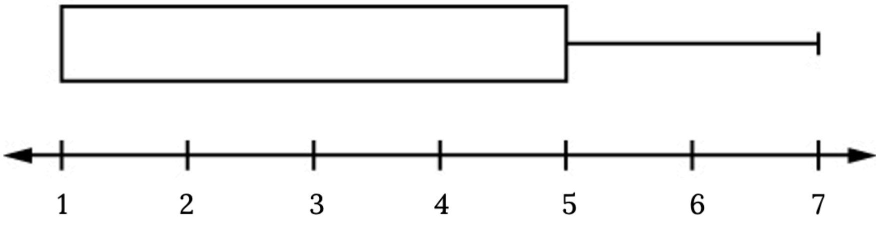


Figure 2.44: Box plot with the same values. *Figure description available at the end of the section.*

In this case, at least 25% of the values are equal to 1. Twenty-five percent of the values are between 1 and 5, inclusive. At least 25% of the values are equal to 5. The top 25% of the values fall between 5 and 7, inclusive.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 2.42: Kindred Grey (2020). *Box plot*. CC BY-SA 4.0. Adaptation of Figure 2.11 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/2-4-box-plots

Figure 2.43: Kindred Grey (2020). *Student heights (box plot)*. CC BY-SA 4.0. Adaptation of Figure 2.12 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/2-4-box-plots

Figure 2.44: Kindred Grey (2020). *Box plot with the same values*. CC BY-SA 4.0. Adaptation of Figure 2.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/2-4-box-plots

Figure Descriptions

Figure 2.42: Horizontal box plot's first whisker extends from the smallest value, one, to the first quartile, two, the box begins at the first quartile and extends to the third quartile, nine, a vertical dashed line is drawn at the median, seven, and the second whisker extends from the third quartile to the largest value of 11.5.

Figure 2.43: Horizontal box plot with first whisker extending from smallest value, 59, to Q1, 64.5, box beginning from Q1 to Q3, 70, median dashed line at Q2, 66, and second whisker extending from Q3 to largest value, 77.

Figure 2.44: Horizontal box plot box begins at the smallest value and Q1, one, until the Q3 and median, five, no median line is designated, and has its lone whisker extending from the Q3 to the largest value, seven.

2.6 Measures of Center

Let's keep working through the acronym SOCS for describing key aspects of our data, this time focusing on the center.

- Shape
- Outliers
- **Center**
- Spread

The "center" is a way of describing the "central tendency" or "typical value" of a dataset. The two most widely used measures of the center of the data are the **mean (average)** and the **median**. Most people are familiar with the ideas of these two: (1) to calculate the mean weight of 50 people, add the 50 weights together and divide by 50, and (2) to find the median weight of the 50 people, order the data and find the number that splits the data into two equal parts.

However, some datasets may be better summarized by one or the other. The most "appropriate" measure of center depends on the shape of the distribution and the presence of extreme values or potential outliers.

The Mean

The mean is the most common measure of the center. The words "mean" and "average" are often used interchangeably. The technical term is "arithmetic mean," and "average" technically refers to a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

When each value in the dataset is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The **sample mean** is denoted by an x with a bar over it, $\overline{x}$, pronounced simply "x bar."

The Greek letter μ (pronounced "mew") represents the **population mean**. We will often use the sample mean to estimate the population mean. One of the requirements for the sample mean to be a good estimate of the population mean is for the sample to be taken truly at random.

Example

Calculate the mean of the sample: 1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4.

Solution
$\bar{x}$ = (1+1+1+2+2+3+4+4+4+4+4)/11 = 2.7

Your Turn!

Calculate the mean of the sample: 7, 10, 14, 14, 15, 21, 38, 38, 38, 56.

Solution
$\bar{x}$ = (7+10+14+14+15+21+38+38+38+56)/10 = 25.1

The Median

The median is generally a better measure of the center when there are extreme values or outliers because it is more **robust**, or not affected by the precise numerical values of those outliers.

Especially for larger datasets, you may choose to use the following location function over the traditional counting method to find the median:

$$L(M) = \frac{n+1}{2}.$$

Remember that this function simply tells you where to look for the median, not the actual value itself, and n is the total number of data values in the sample (sample size).

Once you a have arranged your data in ascending order (smallest to largest), the method of finding your median will differ slightly based on whether you have an odd or even sample size. If n is odd, the median is included in the dataset and is simply the middle value found using the location function. If n is an even number, your location function will give you a decimal value ending in .5, and to find the median, you must calculate the average of the numbers in the $\frac{n}{2}$ and $\frac{n}{2}$ + 1 positions.

For example, if the total number of data values is 97, then $\frac{n+1}{2}$ = $\frac{97+1}{2}$ = 49. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2}$ = $\frac{100+1}{2}$ = 50.5. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are not the same. The upper case letter *M* is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest): 3, 4, 8, 8, 10, 11, 12, 13, 14, 15, 15, 16, 16, 17, 17, 18, 21, 22, 22, 24, 24, 25, 26, 26, 27, 27, 29, 29, 31, 32, 33, 33, 34, 34, 35, 37, 40, 44, 44, 47. Calculate the median.

Solution

To find the median, *M*, first use the formula for the location. The location is:

$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3, 4, 8, 8, 10, 11, 12, 13, 14, 15, 15, 16, 16, 17, 17, 18, 21, 22, 22, 24, 24, 25, 26, 26, 27, 27, 29, 29, 31, 32, 33, 33, 34, 34, 35, 37, 40, 44, 44, 47

$M = \frac{24+24}{2} = 24$

Your Turn!

Calculate the median of the sample: 7, 10, 14, 14, 15, 21, 38, 38, 38, 56.

The Mode

Another measure of the center is the **mode**. The mode is the most frequent value. There can be more than one mode in a dataset as long as those values have the same frequency and that frequency is the highest. A dataset with two modes is called bimodal. For example, if five real estate exam scores are 430, 430, 480, 480, 495, then the dataset is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds during the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

NOTE:

The mode can be calculated for categorical data as well as quantitative data but has different uses and interpretations for each. For example:

- If we had the categorical dataset {red, red, red, green, green, yellow, purple, black, blue}, the mode is red. This is useful to us.
- If we had the quantitative dataset {1.0, 2.1, 2.1, 5.0, 5.1, 5.5, 5.7, 6.1, 6.2, 6.4, 6.6, 7.1, 7.8, 8.1, 8.9}, the numerical mode is 2.1, but that does not do a good job of telling us about the actual **modality**, or where the data is clustered.

Example

Statistics exam scores for 20 students are as follows:

50, 53, 59, 59, 63, 63, 72, 72, 72, 72, 72, 76, 78, 81, 83, 84, 84, 84, 90, 93

Find the mode.

Solution
The most frequent score is 72, which occurs five times. Mode = 72.

Order Relationship of Measures of Center

Consider the following dataset: 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10.

This dataset can be represented by the histogram in Figure 2.45. Each interval has a width of one, and each value is located in the middle of an interval.

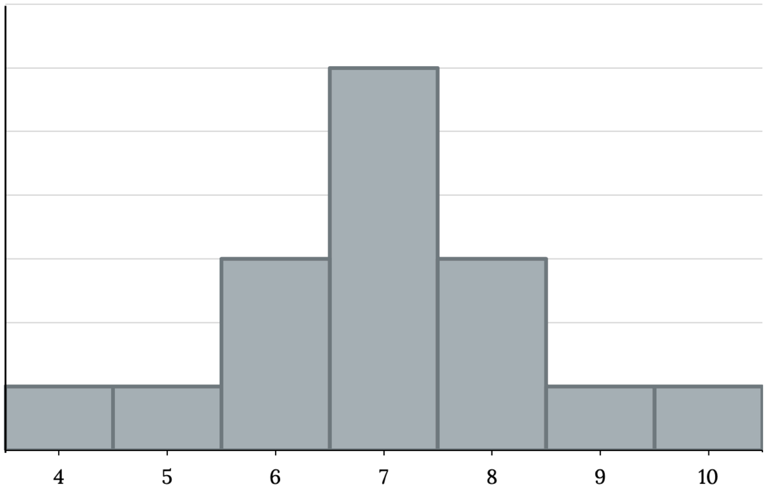


Figure 2.45: Symmetrical distribution. Figure description available at the end of the section.

The histogram displays a symmetrical distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shapes to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. In a perfectly symmetrical distribution, the mean and the median are the same. This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data {4, 5, 6, 6, 6, 7, 7, 7, 7, 8} is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is said to be skewed to the left because it is pulled out to the left.

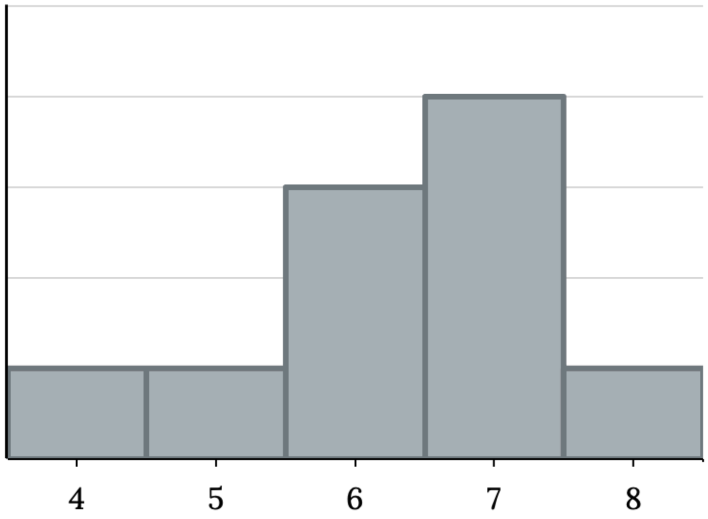


Figure 2.46: Skewed left. Figure description available at the end of the section.

The mean is 6.3, the median is 6.5, and the mode is 7. Notice that the mean is less than the median, and they are both less than the mode. The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data {6, 7, 7, 7, 7, 8, 8, 8, 9, 10} is also not symmetrical. It is skewed to the right.

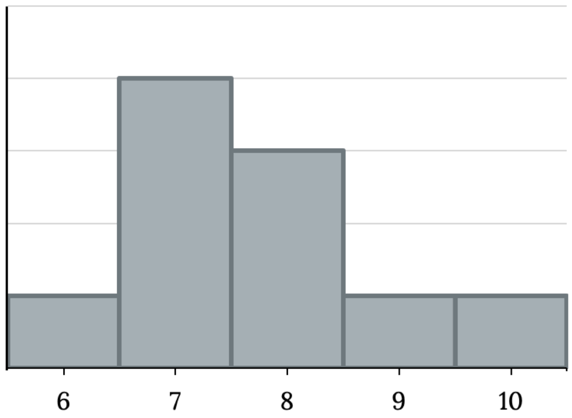


Figure 2.47: Skewed right. Figure description available at the end of the section.

The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, the mean is the largest, while the mode is the smallest. Again, the mean reflects the skewing the most.

To summarize, if the distribution of data is skewed to the left, the mean is often less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

Darnell: 7, 9, 3, 3, 3, 4, 1, 3, 2, 2

Mary: 3, 3, 3, 4, 1, 4, 3, 2, 3, 1

Lee: 2, 3, 4, 4, 4, 6, 6, 6, 8, 3

Make a dot plot for the three authors and compare the shapes.

Darnell's distribution has a right (positive) skew:

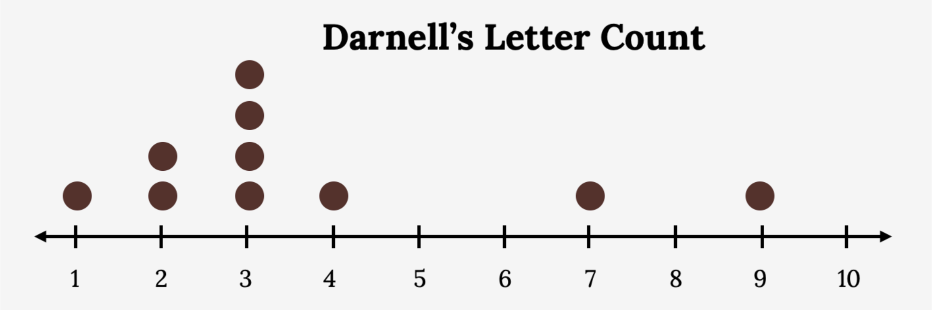


Figure 2.48: Darnell's letter count. Figure description available at the end of the section.

Mary's distribution has a left (negative) skew:

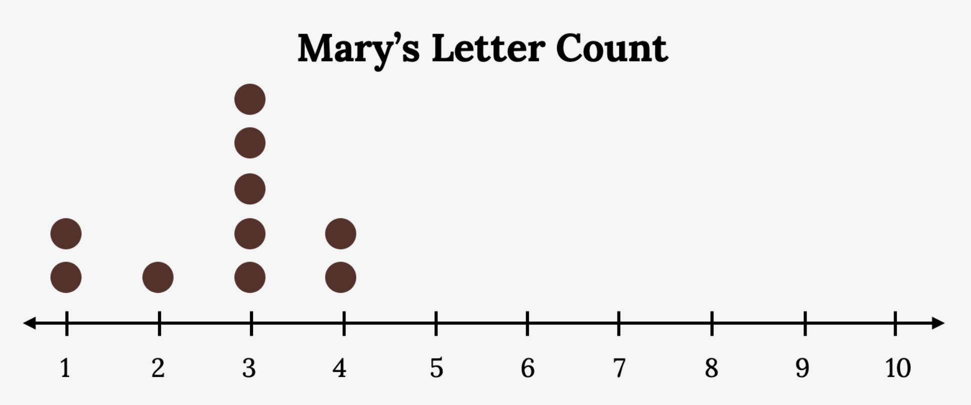


Figure 2.49: Mary's letter count. Figure description available at the end of the section.

Lee's distribution is symmetrically shaped:

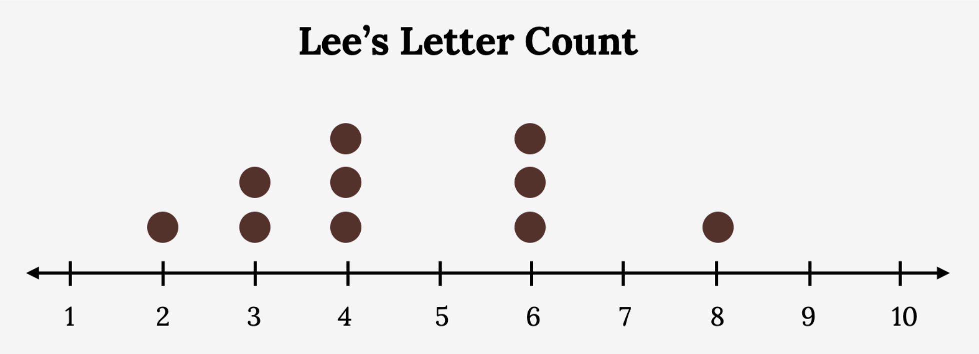


Figure 2.50: Lee's letter count. Figure description available at the end of the section.

Calculate the mean for each.

Solution
Darnell's mean is 3.7. Mary's mean is 2.7. Lee's mean is 4.6.

Calculate the median for each.

Solution
Darnell's median is three. Mary's median is three. Lee's median is four.

Describe any pattern you notice between the shape and the measures of center.

Solution
It appears that the median is closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Your Turn!

Suppose that in a small town of 50 people, one person earns $5,000,000 per year, and the other 49 each earn $30,000. Which is the better measure of the "center," the mean or the median?

Solution
$x = \frac{5{,}000{,}000+49(30{,}000)}{50} = 129{,}400$

$M = 30{,}000$

There are 49 people who earn $30,000 and 1 person who earns $5,000,000. The median is a better measure of the center than the mean because 49 of the values are $30,000 and one is $5,000,000. The $5,000,000 is an outlier. The $30,000 gives us a better sense of the middle of the data.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, we do not know the individual data values (only the intervals and interval frequencies); therefore, we cannot compute an exact mean for the dataset. What we must do is estimate the actual mean by calculating the mean of a frequency table (a data representation in which grouped data is displayed along with the corresponding frequencies). To calculate the mean from a grouped frequency table, we can apply the basic definition of mean:

$$mean = \frac{data\ sum}{number\ of\ data\ values}.$$

We simply need to modify the definition to fit within the restrictions of a frequency table. Since we do not know the individual data values, we can instead find the midpoint of each interval. The midpoint is:

$$\frac{lower\ boundary+upper\ boundary}{2}.$$

We can now modify the mean definition to be:

$$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f} = \frac{\sum fm}{n}$$

where *f* = the frequency of the interval and *m* = the midpoint of the interval.

Example

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade interval	Number of students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Figure 2.51: Blount's statistics test

Find the midpoints for all intervals.

Solution

Grade interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

Figure 2.52: Midpoint

Calculate the sum of the product of each interval frequency.

Solution

$\sum fm$ = 53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5 (3) + 89.5(4) + 95.5(1) = 1,460.25

Calculate the midpoint.

Solution

$$\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$$

Your Turn!

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours teenagers spend on video games	Number of teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

Figure 2.53: Video game data

What is the best estimate for the mean number of hours spent playing video games?

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 2.45: Kindred Grey (2020). *Symmetrical distribution*. CC BY-SA 4.0.

Figure 2.46: Kindred Grey (2020). *Skewed left*. CC BY-SA 4.0.

Figure 2.47: Kindred Grey (2020). *Skewed right*. CC BY-SA 4.0.

Figure 2.48: Kindred Grey (2020). *Darnell's letter count*. CC BY-SA 4.0. Adaptation of Figure 2.21 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode

Figure 2.49: Kindred Grey (2020). *Mary's letter count*. CC BY-SA 4.0. Adaptation of Figure 2.22 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode

Figure 2.50: Kindred Grey (2020). *Lee's letter count*. CC BY-SA 4.0. Adaptation of Figure 2.23 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode

Figure Descriptions

Figure 2.45: Histogram that matches the supplied data. It consists of seven adjacent bars with the x-axis split into intervals of one from four to 10. The heights of the bars peak in the middle and taper symmetrically to the right and left.

Figure 2.46: Histogram that matches the supplied data. It consists of five adjacent bars with the x-axis split into intervals of one from four to eight. The peak is to the right, and the heights of the bars taper down to the left.

Figure 2.47: Histogram that matches the supplied data. It consists of five adjacent bars with the x-axis split into intervals of one from six to 10. The peak is to the left, and the heights of the bars taper down to the right.

Figure 2.48: Dot plot that matches the supplied data for Darnell. The plot uses a number line from one to 10. It shows one x over one, two x's over two, four x's over three, one x over four, one x over seven, and one x over nine. There are no x's over the numbers five, six, eight, and 10.

Figure 2.49: Dot plot that matches the supplied data for Mary. The plot uses a number line from one to 10. It shows two x's over one, one x over two, five x's over three, and two x's over four. There are no x's over the numbers five, six, seven, eight, nine, and 10.

Figure 2.50: Dot plot that matches the supplied data for Lee. The plot uses a number line from one to 10. It shows one x over two, two x's over three, three x's over four, three x's over six, and one x over eight. There are no x's over the numbers one, five, seven, nine, and 10.

2.7 Measures of Spread

We have made it to spread, the final key aspect of the acronym SOCS.

- Shape
- Outliers
- Center
- **Spread**

A complement to the center of a distribution is the data's **spread** (also known as variation or variability). In some datasets, the values are concentrated closely, while they are more spread out in others. Some rough measures of spread we have already discussed are the range and IQR. The most common measure of spread is the standard deviation.

Similar to measures of center, the shape of the distribution and the presence of extreme values can dictate what measure of spread is most appropriate to describe the distribution.

The Interquartile Range

Recall the interquartile range (IQR):

IQR = $Q_3 - Q_1$.

In addition to helping us establish our fences and identify outliers, the IQR indicates the spread of the middle half (middle 50%) of the data. The IQR can be used as a somewhat rough but very robust measure of spread when outliers may be present. It is often used alongside the median to describe the center and spread of skewed distributions.

Simply showing the five-number summary or a box plot can be a good way to get all of the information for a skewed dataset in one place.

The Standard Deviation

The **standard deviation** is a measure of spread that assesses how dispersed values are from their mean. It is essentially the "average" deviation—the distance of each observation from the mean.

Not only does it provide a numerical measure of the overall amount of variation in a dataset, it can also be used for other purposes.

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

By extension, s^2 represents the sample **variance**, and the lowercase Greek letter σ^2 represents the population variance.

The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation. It must always greater than or equal to zero.

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

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

Calculating the Standard Deviation

The procedure to calculate the standard deviation can be tedious and depends on whether the data are from the entire population or a sample. The calculations are similar but not identical.

If x is a number, then the difference "x – mean" is its deviation. In a dataset, there are as many deviations as there are items in the set. The deviations can show how spread out the data are from the mean. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. If the numbers belong to a population, a deviation is $x - \mu$ in symbols. For sample data, a deviation is $x - \overline{x}$ in symbols. If you add the deviations, the sum is always zero, so you cannot simply add the deviations to get the spread of the data. You can fix this by squaring the deviations, making them positive numbers; therefore, the sum will also be positive.

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

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $n - 1$, one less than the number of items in the sample. Why not divide by n for a sample? The answer has to do with the population variance. The sample variance is an estimate of the population variance. Based on the theoretical mathematics underlying these calculations, dividing by $n - 1$ gives a better estimate of the population variance.

Formulas

The sample standard deviation

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

The population standard deviation

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

NOTES:

- The variance, whether population (σ^2) or sample (s^2), can be obtained if you do not apply the square root in their respective formulas
- Though we typically rely on technology to calculate the standard deviation in practice, please note:
  - In the sample standard deviation formula, the denominator is $n - 1$.
  - In the population standard deviation formula, the denominator is N.
  - You may need to indicate on your technology of choice which form of the formula you want to use.
- We will often use the sample standard deviation or variance to estimate the population standard deviation or variance.

Example

The teacher of a fifth grade class was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a sample of $n = 20$ fifth grade students. The ages are rounded to the nearest half year: 9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5.

First, try to find the mean and standard deviation by hand. Here is a table with the intermediate steps:

X	Deviations	Deviations^2
9	9 – 10.525 = –1.525	$(-1.525)^2$ = 2.325625
9.5	9.5 – 10.525 = –1.025	$(-1.025)^2$ = 1.050625
9.5	9.5 – 10.525 = –1.025	$(-1.025)^2$ = 1.050625
10	10 – 10.525 = –0.525	$(-0.525)^2$ = 0.275625
10	10 – 10.525 = –0.525	$(-0.525)^2$ = 0.275625
10	10 – 10.525 = –0.525	$(-0.525)^2$ = 0.275625
10	10 – 10.525 = –0.525	$(-0.525)^2$ = 0.275625
10.5	10.5 – 10.525 = –0.025	$(-0.025)^2$ = 0.000625
10.5	10.5 – 10.525 = –0.025	$(-0.025)^2$ = 0.000625
10.5	10.5 – 10.525 = –0.025	$(-0.025)^2$ = 0.000625
10.5	10.5 – 10.525 = –0.025	$(-0.025)^2$ = 0.000625
11	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625
11	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625
11	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625
11	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625
11	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625
11	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625
11.5	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625
11.5	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625
11.5	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625
-	-	**Total = 9.7375**

Figure 2.54: Fifth grade ages

Verify your answers with your choice of technology.

Solution

Mean = $\bar{x}$ = $\frac{9+9.5+9.5+10+10+10+10+10.5+10.5+10.5+10.5+11+11+11+11+11+11+11.5+11.5+11.5}{20}$

= $\frac{210.5}{20}$ = 10.525

The variance may be calculated by hand according to the table above.

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

s^2 = $\frac{9.7375}{20-1}$ = 0.5125. Notice that instead of dividing by n = 20, the calculation divided by n – 1 = 20 – 1 = 19 because the data is a sample.

The sample standard deviation, s, is equal to the square root of the sample variance:

s = $\sqrt{0.5125}$ = 0.715891 which is rounded to two decimal places, s = 0.72.

Your Turn!

On a baseball team, the ages of each of the players are as follows:

21, 21, 22, 23, 24, 24, 25, 25, 28, 29, 29, 31, 32, 33, 33, 34, 35, 36, 36, 36, 36, 38, 38, 38, 40

First, try to find the mean and standard deviation by hand. If you get stuck or want to check your work, plug it into your calculator or use your computer software.

The standard deviation, s or σ, is either zero or larger than zero. Describing the data with reference to the spread is called "variability." The variability in data depends upon the method by which the outcomes are obtained (e.g., by measuring or random sampling). When the standard deviation is zero, there is no spread; all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and it is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out from the mean; outliers can make s or σ very large.

The Standard Deviation in Context

When first presented, the standard deviation can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that the standard deviation can be very helpful in symmetrical distributions, but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, always graph your data, displaying it in a histogram or a box plot.

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

If one were also part of the dataset, then 1 is two standard deviations to the left of 5 because 5 + (−2)(2) = 1.

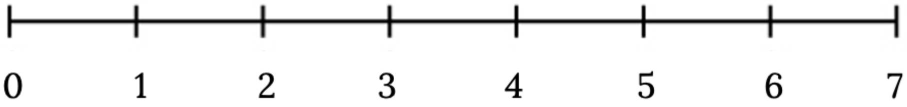


Figure 2.55: Number line. Figure description available at the end of the section.

- In general, a value = mean + (#ofSTDEV)(standard deviation)
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer.
- 1 is two standard deviations less than the mean of 5 because 1 = 5 + (−2)(2).

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

Sample:

$$x = \overline{x} + (\#ofSTDEVs)(s)$$

Population:

$$x = \mu + (\#ofSTDEVs)(\sigma)$$

Example

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

Rosa waits for seven minutes:

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

Binh waits for one minute:

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is four minutes less than the average of five minutes.
- Binh's wait time of one minute is two standard deviations below the average of five minutes.

Your Turn!

Recall the previous example about the age of fifth grade students where $\overline{x}$ = 10.525 and s^2 = 0.7159.

Find the value that is one standard deviation above the mean. Find ($\overline{x}$ + 1s).

Solution
$\overline{x}$ + 1s = 10.53 + (1)(0.72) = 11.25

Find the value that is two standard deviations below the mean. Find ($\overline{x}$ – 2s).

Solution
$\overline{x}$ – 2s = 10.53 – (2)(0.72) = 9.09

Find the values that are 1.5 standard deviations from (below and above) the mean.

Solution
$\overline{x}$ – 1.5s = 10.53 – (1.5)(0.72) = 9.45

$\overline{x}$ + 1.5s = 10.53 + (1.5)(0.72) = 11.61

z-Scores

The standard deviation can also be used to calculate a measure of location called a **z-score**. It represents the number of standard deviations between a given observation and its mean (#ofSTDEVs above), which is often denoted with just the letter *z*. In symbols, the formulas become:

z-Score formulas		
Sample	$x = \overline{x} + zs$	$z = \frac{x-\overline{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x-\mu}{\sigma}$

Figure 2.56: z-Score formulas

Not only are *z*-scores a useful measure of location for specific observations, they can also be used for other purposes. If two datasets have different means and standard deviations, then comparing the data values directly can be misleading. However, using *z*-scores, it is possible to put things on a level playing field to compare them.

- For each data value, calculate the number of standard deviations between the value and its mean.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.

- #ofSTDEVs = $\frac{value - mean}{standard\ deviation}$
- Compare the results of this calculation.

To understand the concept, suppose X ~ N(5, 6) represents weight gains for one group of people who are trying to gain weight in a six-week period, and Y ~ N(2, 1) measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since x = 17 and y = 4 are each two standard deviations to the right of their means, they represent the same, standardized weight gain relative to their means.

Example

John and Ali, two students from different high schools, wanted to find out who had the highest GPA when compared to the rest of his school. Which student had the highest GPA when compared to his school?

Student	GPA	School mean GPA	School standard deviation
John	2.85	3.0	0.7
Ali	77	80	10

Figure 2.57: GPA comparisons

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \text{\#ofSTDEVs} = \frac{value - mean}{standard\ deviation} = \frac{x - \mu}{\sigma}$$

Solution

For John, z = #ofSTDEVs = $\frac{2.85 - 3}{0.7}$ = -0.21

For Ali, z = #ofSTDEVs = $\frac{77 - 80}{10}$ = -0.3

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** her school's mean.

John's z-score of −0.21 is higher than Ali's z-score of −0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Your Turn!

Angie and Beth, two swimmers from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to the rest of her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team mean time	Team standard deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Figure 2.58: Swim time comparisons

Identifying "Unusual" Observations

Recall we have already established our fence rules for numerically identifying outliers in any distribution. However, for most symmetric and bell-shaped distributions, anything outside of two standard deviations (a z-score below -2 or greater than 2) is considered "unusual". We will learn more about this in later chapters, but generally an observation should be within ±2 standard deviations 95% of the time. However, considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule.

Example

The distribution of heights for US males is considered to be symmetric and bell-shaped, with an average of 69.7 inches and a 2.8 inch standard deviation. How tall would a male have to be to be considered "unusually" tall in the US?

Solution
69.7 + (2*2.8) = 75.3 inches

Your Turn!

The distribution of heights for US females is considered to be symmetric and bell-shaped, with an average of 64.4 and a 2.4 inch standard deviation. How short would a female have to be to be considered “unusually” short in the US?

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 2.55: Kindred Grey (2020). *Number line*. CC BY-NC 4.0.

Figure Descriptions

Figure 2.55: Blank number line in intervals of one from zero to seven.

Chapter 2 Wrap-Up

Concept Check

Take this quiz to check your comprehension of this chapter.

If you are using an offline version of this text, access the quiz for this chapter via the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=120#h5p-72

Section Resources

If you are using an offline version of this text, access these materials by visiting https://doi.org/10.7294/26207456.

2.1 Descriptive Statistics and Frequency Distributions

2.2 Displaying and Describing Categorical Distributions

2.3 Displaying Quantitative Distributions

2.4 Describing Quantitative Distributions

2.5 Measures of Location and Outliers

2.6 Measures of Center

2.7 Measures of Spread

Key Terms

Try to define the terms below on your own. Check your response by clicking on the term, or looking at the end-of-book glossary!

2.1 Descriptive Statistics and Frequency Distributions

- **Descriptive statistics**
- **Graphical descriptive methods**
- **Numerical descriptive methods**
- **Distribution**
- **Frequency**
- **Lower class limit**
- **Upper class limit**
- **Class width**
- **Class midpoint**
- **Relative frequency**
- **Cumulative relative frequency**

2.2 Displaying and Describing Categorical Distributions

- **Categorical data**
- **Mode**
- **Variability**

2.3 Displaying Quantitative Distributions

- **Quantitative data**
- **Discrete data**
- **Ordinal categorical data**

2.4 Describing Quantitative Distributions

- **Shape**
- **Outliers**
- **Center**
- **Spread (variation, variability)**
- **Modality**

2.5 Measures of Location and Outliers

- **Median**

2.6 Measures of Center

- **Mean (average)**
- **Sample mean**
- **Population mean**
- **Robust**

2.7 Measures of Spread

- **Standard deviation**
- **Sample**
- **Population**
- **Variance**
- ***z*-score**

Extra Practice

Extra practice problems are available at the end of the book (Chapter 2 Extra Practice).

CHAPTER 3: BIVARIATE DESCRIPTIVE STATISTICS

3.1 Introduction to Bivariate Data

Learning Objectives

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

- Display and describe relationships in bivariate data (categorical and quantitative)
- Describe bivariate quantitative data numerically
- Understand and apply the ideas of simple linear regression

Professionals often want to know how two (or more) variables are related. For example, is there a relationship between a student's grade on their second math exam and their grade on the final? If there is a relationship, what is the relationship and how strong is it?

In another example related to Figure 3.1, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

Figure 3.1: Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. Figure description available at the end of the section.

The type of data described in these examples is bivariate data ("bi" for two variables). We could have:

- A categorical variable vs. another categorical variable
- A categorical variable vs. a quantitative variable
- A quantitative vs. a quantitative variable

This section will briefly discuss displaying a quantitative variable with a categorical grouping variable and then focus on displaying two categorical variables. The rest of this chapter will then focus on relationships between two quantitative variables.

Picturing Bivariate Variables

When it comes to displaying a quantitative variable as a response vs. a categorical variable as a predictor, the methods we will discuss mainly apply to situations where we have a quantitative response variable being measured and want to further break it down by another categorical grouping variable. Some methods are simply an overlaid line graph or histogram.

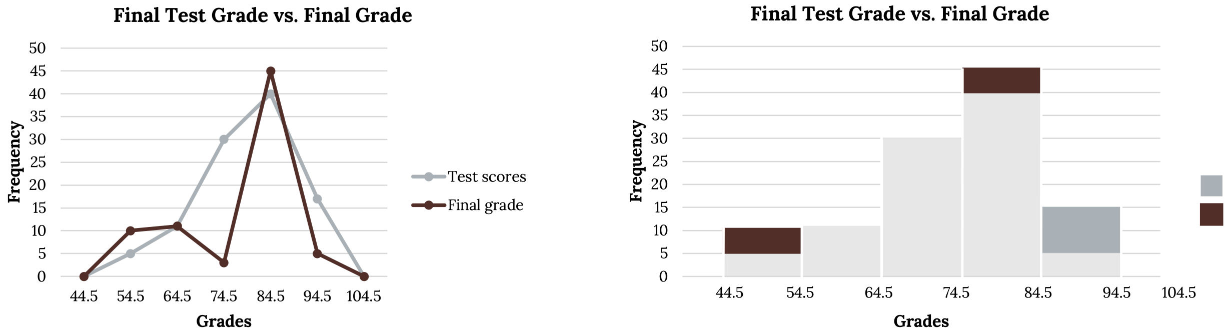


Figure 3.2: Line graph and histogram. [Figure description available at the end of the section](#).

The above options may work well in some cases, like when the bins for each group line up well. For most cases, however, a better option can often be a comparative box plot:

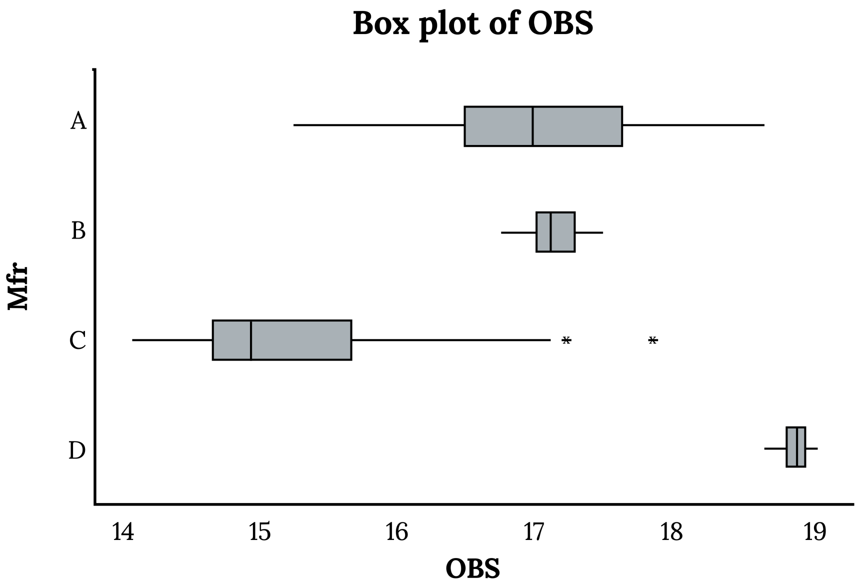


Figure 3.3: Comparative box plot. [Figure description available at the end of the section](#).

Heat maps are particularly well suited to handle situations where there is a geographical or spatial element.

Figure 3.4: Heat map. Figure description available at the end of the section.

There are numerical methods to further analyze categorical response and quantitative predictor variables, but they get pretty complicated mathematically and are beyond the scope of this course.

Picturing Bivariate Categorical Variables

We will begin by examining the relationship between two categorical variables visually. The options below build off some ideas we have discussed in relation to univariate categorical data.

- Univariate frequency tables → Contingency tables
- Univariate bar chart → Stacked or grouped bar chart

Contingency Tables

A **contingency table** portrays data in a way that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, will revisit contingency tables and use them in another manner.

Example

Suppose a study of speeding violations and drivers who use cell phones produced the following data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Uses cell phone while driving	25	280	305
Does not use cell phone while driving	45	405	450
Total	70	685	755

Figure 3.5: Driving violations

The total number of people in the sample is 755. The marginal row totals are 305 and 450, and the marginal column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Your Turn!

The figure below contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the US.

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

Figure 3.6: US crime index rates

Find the following:

1. Marginal frequencies
2. Overall total
3. Marginal relative frequencies
4. Conditional percentages of type of crime in each given year

Variations on Bar Charts

The following variations on bar charts can also help us see relationships between two categorical variables, providing us with a little more visual information than a contingency table:

- Stacked bar charts
- Grouped or side-by-side bar charts

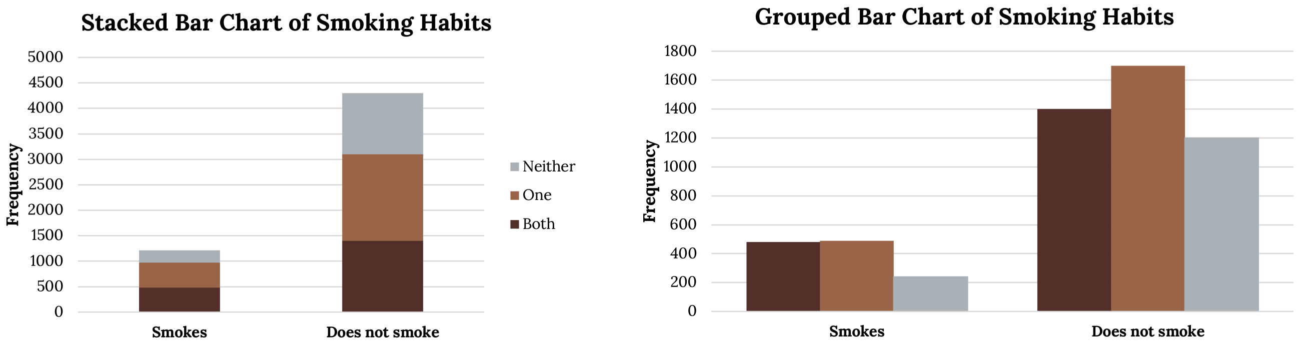


Figure 3.7: Stacked bar chart and grouped bar chart. *Figure description available at the end of the section.*

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 3.1: Aaron Huber (2018). *Man holding engines*. Unsplash license. https://unsplash.com/photos/man-holding-engines-KxeFuXta4SE

Figure 3.2: Kindred Grey (2024). *Line graph and histogram*. CC BY 4.0.

Figure 3.3: Kindred Grey (2024). *Comparative box plot*. CC BY 4.0.

Figure 3.4: Clay Banks (2020). *Red and Black Heart Illustration*. Unsplash license. https://unsplash.com/photos/red-and-black-heart-illustration-U0-r0JMypE0

Figure 3.7: Kindred Grey (2024). *Stacked bar chart and grouped bar chart*. CC BY 4.0.

Figure Descriptions

Figure 3.1: Man inspecting an engine in an auto shop.

Figure 3.2: Left: Two lines (one for test scores and one for final grades) connected by points. Both peak around 84.5 grade with a frequency of 45. Right: boxes on a graph next to each other. Three of the five have extra boxes stacked on top of one another, indicating that the values for test scores and final grades are different from one another for these three frequencies.

Figure 3.3: Four box plots of varying widths, medians, and outliers.

Figure 3.4: Map of the world with orange circles varying in size placed on the map and overlap

Figure 3.7: Left: stacked bar chart with neither, one, and both represented in different colors stacked in the same bar labeled “smokes”. There is another bar labeled “does not smoke” with the same three categories stacked on top of one another. Right: Smokes category is on the left, but this time with neither, one, and both columns placed side by side. Same for “does not smoke”.

3.2 Visualizing Bivariate Quantitative Data

Bivariate Quantitative Data

When we are looking at **bivariate data**, we first need to decide if changing one variable seems to lead to a change in the other. A **response variable** (also called y, dependent variable, and predicted variable) measures or records an outcome of a study. An **explanatory variable** (also called x, independent variable, and predictor variable) explains changes in the response variable.

In the rest of this chapter, we will be studying "simple linear regression." Note that this does not imply that these ideas are "simple" but just that we are working with one independent variable (x) and a linear relationship. This involves data that fits a line in two dimensions.

When considering the relationship between two quantitative variables:

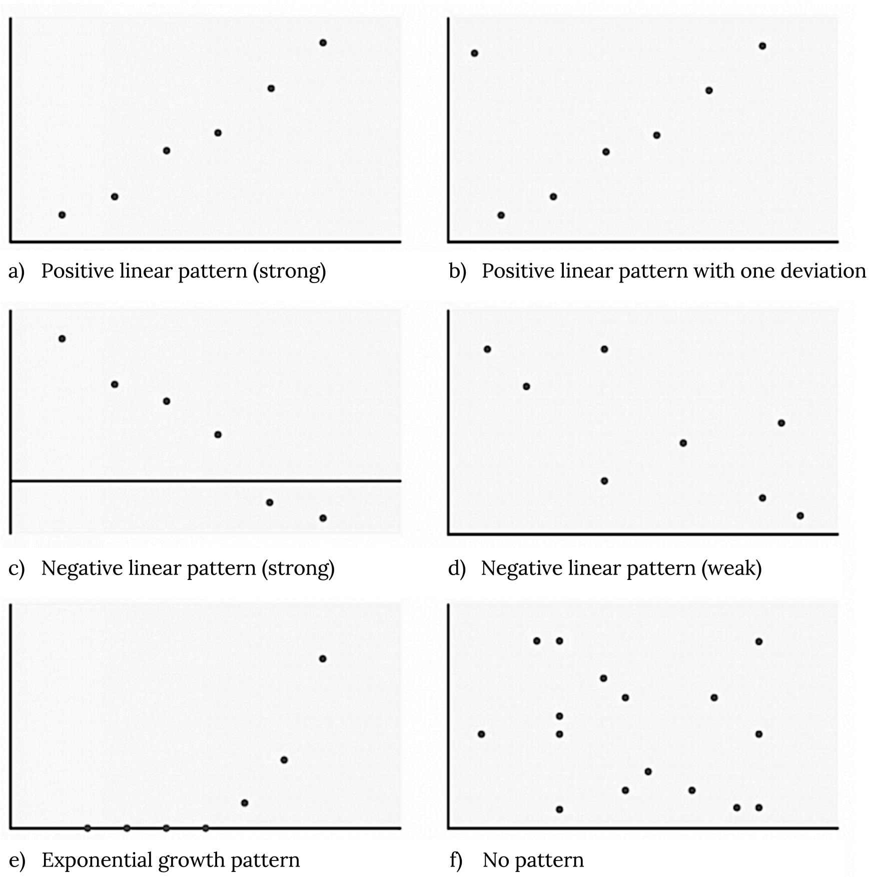
1. Start with a graph (scatter plot).
2. Look for an overall pattern and deviations from the pattern.
3. Use numerical descriptions of the data and overall pattern (correlation, coefficient of determination).
4. Consider a mathematical model (regression).

Scatter Plots

Before we discuss linear regression and correlation, we need to examine a way to display the relation between the variables x and y. The most common and easiest way is a scatter plot. A scatter plot shows a lot about the relationship between the variables. When you look at a scatter plot, you want to notice the overall pattern and any potential deviations from the pattern. You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are together. When looking at a scatter plot you always want to note:

- Shape
- Trend
- Strength

The following scatter plot examples illustrate these concepts.



a) Positive linear pattern (strong)

b) Positive linear pattern with one deviation

c) Negative linear pattern (strong)

d) Negative linear pattern (weak)

e) Exponential growth pattern

f) No pattern

Figure 3.8: Scatter plot configurations. *[Figure description available at the end of the section](#)*.

Shape

Although we may see other shapes in a scatter plot, we are currently only interested in applying these ideas when we see a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line, which indicates no relationship. If we think that the points show a linear relationship, we draw a line on the scatter plot. Later, we will learn to calculate this line through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable.

Trend

If we do see a linear pattern, what sort of relationship is there? A positive trend is seen when increasing x also increases y. On the other hand, a negative (inverse) trend is seen when increasing x appears to cause y to decrease. In other words:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable
- High values of one variable occurring with low values of the other variable

Strength

At this point, we can think about the strength of a relationship by asking how tightly the points on a scatter plot fit the linear pattern. A stronger relationship has points clustered together closely, while in a weaker one, points are more spread out. The strength of a relationship is not always apparent in a scatter plot, but we will see them measured numerically in the future.

Example

Does the scatter plot appear linear? Strong or weak? Positive or negative?

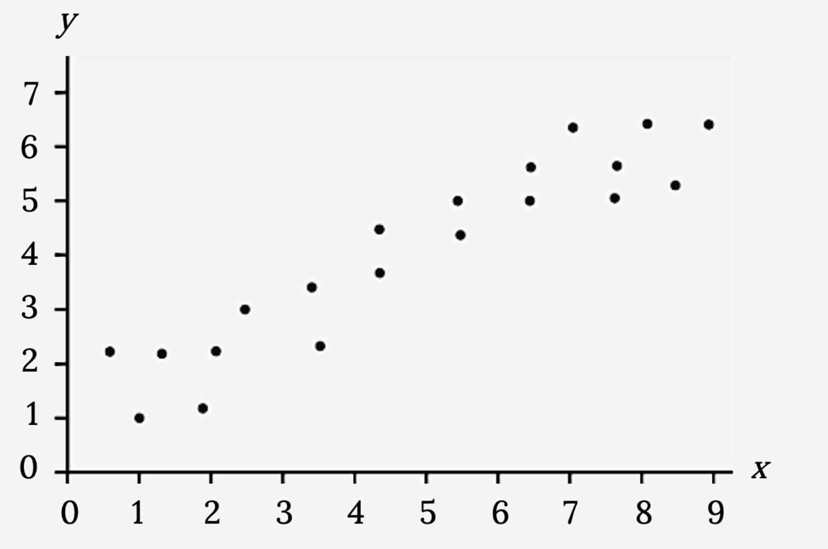


Figure 3.9: Scatter plot 1. Figure description available at the end of the section.

Solution
The data appear to be linear with a strong, positive correlation.

Does the scatter plot appear linear? Strong or weak? Positive or negative?

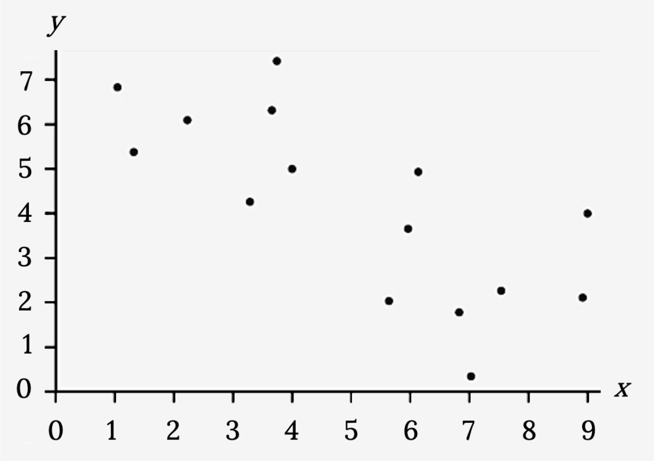


Figure 3.10: Scatter plot 2. Figure description available at the end of the section.

Solution

The data appear to be linear with a weak, negative correlation.

Does the scatter plot appear linear? Strong or weak? Positive or negative?

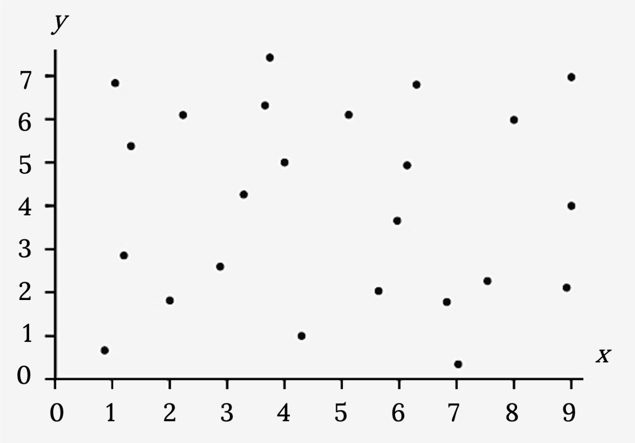


Figure 3.11: Scatter plot 3. Figure description available at the end of the section.

Solution

The data appear to have no correlation.

Your Turn!

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game seems to go up in response to the number of hours she practices her jump shot each week. She records the following data:

Hours practicing jump shot (x)	Points scored in a game (y)
5	15
7	22
9	28
10	31
11	33
12	36

Figure 3.12: Amelia's points

Construct a scatter plot, and state whether Amelia's hypothesis appears to be true.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 3.8: Kindred Grey (2020). *Scatter plot configurations*. CC BY-SA 4.0. Adaptation of Figures 12.6, 12.7, and 12.8 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-2-scatter-plots

Figure 3.9: Kindred Grey (2020). *Scatter plot 1*. CC BY-SA 4.0. Adaptation of Figure 12.26 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-practice

Figure 3.10: Kindred Grey (2020). *Scatter plot 2*. CC BY-SA 4.0. Adaptation of Figure 12.27 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-practice

Figure 3.11: Kindred Grey (2020). *Scatter plot* 3. CC BY-SA 4.0. Adaptation of Figure 12.28 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-practice

Figure Descriptions

Figure 3.8: Six scatterplots showing different patterns. First: positive linear pattern (strong)—shows dots in an almost perfect line from bottom left of graph to top right. Second: linear pattern with one deviation—shows the same pattern as first scatterplot with one outlier in the top left corner. Third: negative linear pattern (strong)—shows dots in an almost perfect line from top left to bottom right of graph. Fourth: negative linear pattern (weak)—shows dots from top left to bottom right of graph nowhere near a perfect line, but not completely random. Fifth: exponential growth pattern—shows a few dots on the x axis from left to right in a horizontal line and then gradually the dots move upwards towards the top right corner creating an upwards curve. Sixth: no pattern—random dots all over the graph.

Figure 3.9: Scatterplot with several points plotted in the first quadrant. The points form a clear pattern, moving upward to the right. The points do not line up , but the overall pattern can be modeled with a line.

Figure 3.10: Scatterplot with several points plotted in the first quadrant. The points move downward to the right. The overall pattern can be modeled with a line, but the points are widely scattered.

Figure 3.11: Scatter plot with several points plotted all over the first quadrant. There is no pattern.

3.3 Measures of Association

You can look at the scatter plot and see that a linear relationship seems reasonable, and you can identify a positive or negative trend, but how can you tell more about this relationship? While it is always good practice to first examine things visually, you may find that deciphering a scatter plot can be tricky, especially when it comes to the strength of a relationship. The next step is then to calculate numerical measures of this association.

The Correlation Coefficient, *r*

The **correlation coefficient**, r, developed by Karl Pearson in the early 1900s, is a numerical measure of the strength and direction of the linear association between the independent variable x and the dependent variable y.

The correlation coefficient can be calculated using the formula:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where n = the number of data points.

The formula for r is formidable, so I would not recommend doing this by hand, but technology can make quick work of the calculation.

If you suspect a linear relationship between x and y, then r can measure the strength of the linear relationship.

What the VALUE of r tells us:

- The value of r is always between −1 and +1 (i.e., $-1 \leq r \leq 1$).
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to −1 or to +1 indicate a stronger linear relationship between x and y.
- If $r = 0$, there is likely no linear correlation. It is important to view the scatter plot, however, because data exhibiting a curved or horizontal pattern may have a correlation of 0.
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us:

- A positive value of r means that when x increases, y tends to increase, and when x decreases, y tends to decrease (positive correlation).
- A negative value of r means that when x increases, y tends to decrease, and when x decreases, y tends to increase (negative correlation).
- The sign of r is the same as the sign of the slope of the best-fit line (b).

NOTE:

Strong correlation does not suggest that x causes y or y causes x. We say "correlation does not imply causation."

Example

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200.

Third exam score (x)	Final exam score (y)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Figure 3.13: Third and final exam scores data

A scatter plot showing the scores on the final exam based on scores from the third exam is shown below.

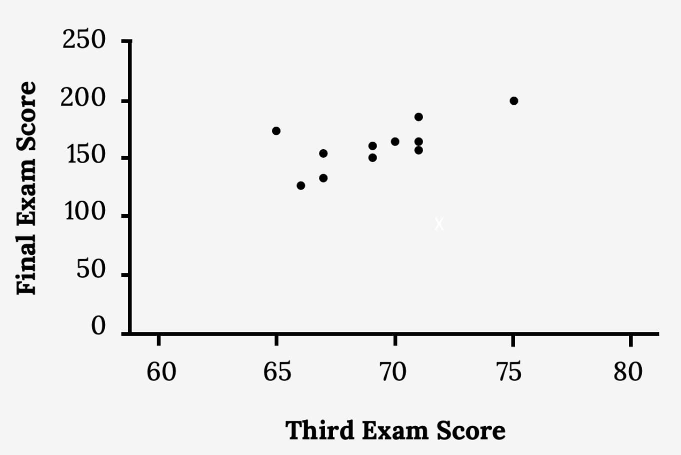


Figure 3.14: Third and final exam scores scatter plot. [Figure description available at the end of the section](#).

Find the correlation coefficient.

Solution

Using technology we would find the correlation coefficient is $r = 0.6631$.

Your Turn!

Match the following scatter plots with the three descriptions of correlation coefficients below them.

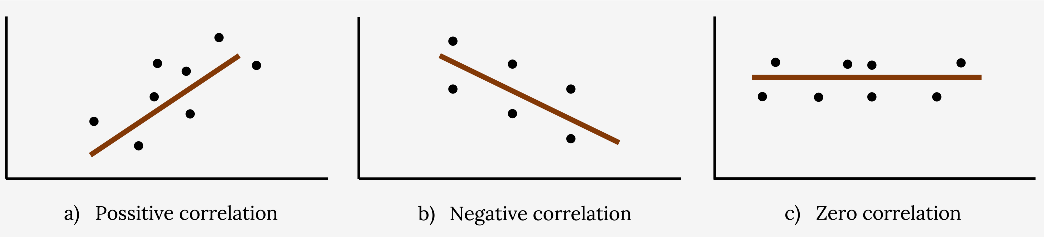


Figure 3.15: Matching scatter plots to correlation coefficients. [Figure description available at the end of the section](#).

- $-1 < r < 0$
- $r = 0$
- $0 < r < 1$

Solution
(a) $0 < r < 1$, (b) $-1 < r < 0$, (c) $r = 0$

The Coefficient of Determination, r^2

While the **coefficient of determination** (r^2) is (obviously) the square of the correlation coefficient, it is usually stated as a percent rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Example

Recall our previous example using a student's third exam scores to predict their final exam scores, in which the correlation coefficient is $r = 0.6631$.

Find the coefficient of determination.

Solution
$r^2 = 0.66312^2 = 0.4397$

Interpret of r^2 in the context of this example.

Solution
Approximately 44% of the variation (0.4397 is approximately 0.44) in the final exam scores can be explained by the variation in the grades on the third exam, using the best-fit regression line.

Therefore, approximately 56% of the variation (1 – 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the scores on the third exam, using the best-fit regression line. This is seen as the scattering of the points about the line.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 3.14: Kindred Grey (2020). *Third and final exam scores scatter plot.* CC BY-SA 4.0. Adaptation of Figure 12.9 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation

Figure 3.15: Kindred Grey (2020). *Matching scatter plots to correlation coefficients.* CC BY-SA 4.0. Adaptation of Figure 12.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation

Figure Descriptions

Figure 3.14: Scatter plot of the data provided. The third exam score is plotted on the x-axis, and the final exam score is plotted on the y-axis. The points form a strong, positive, linear pattern.

Figure 3.15: Three scatter plots with lines of best fit. The first scatterplot shows points ascending from the lower left to the upper right. The line of best fit has positive slope. The second scatter plot shows points descending from the upper left to the lower right. The line of best fit has negative slope. The third scatter plot of points form a horizontal pattern. The line of best fit is a horizontal line.

3.4 Modeling Linear Relationships

If you knew the length of someone's pinky (smallest finger), do you think you could predict that person's height? Imagine collecting data on this and constructing a scatter plot of the points. Then draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slope and the y-intercept, write your equation of "best fit." According to your equation, what is the predicted height for a pinky length of 2.5 inches? You have just started the process of linear regression.

Linear Regression

Data rarely perfectly fit a straight line, but we can be satisfied with rough predictions. Typically, if a dataset has a scatter plot that appears to "fit" a straight line called a line of best fit or least-squares line. This process of fitting the best-fit line is called **linear regression**.

The equation of the regression line is $\hat{y}$ = a + bx.

The $\hat{y}$ is read "y hat" and is the estimated value of y obtained using the regression line. It may or may not be equal to values of y observed from the data.

The sample means of the x values and the y values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point ($\bar{x}$, $\bar{y}$).

The **slope**, b, can be written as $b = r\left(\frac{s_y}{s_x}\right)$, where s_y is the standard deviation of the y values and s_x is the standard deviation of the x values. Note that the slope is directly calculated using r, the correlation coefficient, discussed in previous sections.

The **y-intercept**, a, can then be calculated by using the slope, and means of x and y.

Example

Recall our previous example:

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Third exam score (x)	Final exam score (y)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Figure 3.16: Third and final exam scores data

Solution
We have found:

- An apparent linear relationship in the scatterplot
- The correlation coefficient is r = 0.6631
- The coefficient of determination is r^2 = 0.66312 = 0.4397

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best fits the data. If you were to fit a line by eye, you may draw different lines. We most often use what is called a least-squares regression line to obtain the best fit line. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the vertical distance of each point to the line is made as small as possible. This best fit line is called the **least-squares regression line**.

Consider the following diagram. Each point of data is of the form (x, y), and each point of the line of best fit using least-squares linear regression has the form (x, $\hat{y}$).

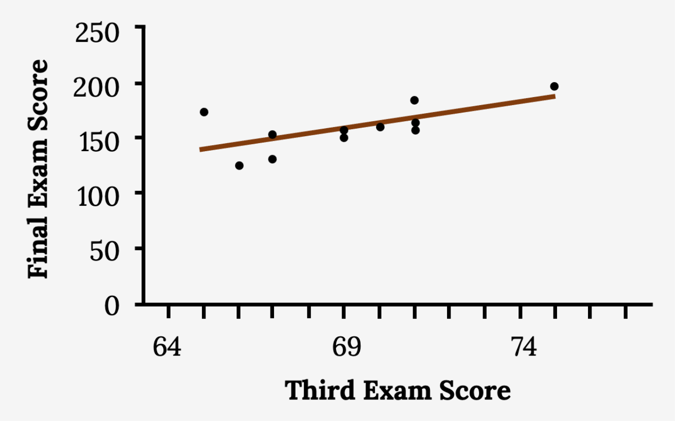


Figure 3.17: Line of best fit. Figure description available at the end of the section.

Solution

The line of best fit is: ŷ = −173.51 + 4.83*x*

Your Turn!

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the figure below show different depths' maximum dive times in minutes. Use your calculator to find the least-squares regression line and predict the maximum dive time for 110 feet.

Depth (*x*) (in feet)	Maximum dive time (*y*) (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

Figure 3.18: SCUBA diver stats

Understanding Slope

The **slope** of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

Interpretation: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable on average.

Example

[Previous Example Continued]

The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points on average.

Understanding the *y*-Intercept

The ***y*-intercept** of the line, a, can tell us what we would predict the value of y to be when x is 0. This may make sense in some cases, but in many, it may not make sense for x to be equal to 0, therefore the y-intercept may not be useful.

Example

[Previous Example Continued]

The y-intercept of the line is -173.51.

Interpretation: In this context, it does not really make sense for x to be 0 (unless a student did not take the exam or try at all). Therefore our y-intercept does not make sense.

Prediction

The next and most useful step in regression is to actually use that equation to predict future values of y.

Recall our example in which we examined the scatter plot and found the **correlation coefficient** and **coefficient of determination**. We found the equation of the best-fit line for the final exam grade as a function of the grade on the third exam. We can now use the least-squares regression line for prediction.

Example

[Previous Example Continued]

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received a score of 73 on the third exam. The exam scores (x values) range from 65 to 75. Since 73 is between the x values 65 and 75, substitute x = 73 into the equation. Then:

$$y = -173.51 + 4.83\left(73\right) = 179.08$$

Solution
We can predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution
145.27

What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution
The x values in the data are between 65 and 75. Ninety is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. Even though it is possible to enter 90 into the equation for x and calculate a corresponding y value, the y value that you get will not be reliable.

Your Turn!

Data is collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$$\hat{y} = 72.5 + 2.8x$$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 3.17: Kindred Grey (2020). *Line of best fit.* CC BY-SA 4.0. Adaptation of Figure 12.11 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation

Figure Descriptions

Figure 3.17: Scatter plot of exam scores with a line of best fit. Weak positive correlation.

3.5 Cautions about Regression

While regression is a very useful and powerful tool, it is also commonly misused. The main things we need to keep in mind when interpreting our results are:

1. Linearity
2. Correlation (or association) does not imply causation
3. Extrapolation
4. Outliers and influential points

Linearity

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use the methods we are discussing.

Correlation Does Not Imply Causation

Even when there is an apparent linear relationship and a reasonable value of r, there can always be confounding or lurking variables at work. Be wary of spurious correlations and make sure the connection you are making makes sense!

There are also often situations where it may not be clear which variables affect each other. Does lack of sleep lead to higher stress levels, or do high stress levels lead to lack of sleep? Which came first, the chicken or the egg? Sometimes these may not be answerable, but at least we are able to show an association there.

Extrapolation

Remember, it is always important to plot a scatter diagram first. If the plot suggests the variables have a linear relationship, then it is reasonable to use a best-fit line to make predictions for y given x within the domain of x values in the sample data, though not necessarily for x values outside that domain. The process of predicting inside of the observed x values observed in the data is called interpolation. The process of predicting outside of the observed x values observed in the data is called **extrapolation**.

Recall our example from the previous section. You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x values in the sample data, which are between 65 and 75.

To understand just how unreliable the prediction can be outside of the observed x values observed in the data, make the substitution $x = 90$ in the equation:

$$y = -173.51 + 4.83\,(90) = 261.19$$

The final exam score is predicted to be 261.19. The largest a final exam score could be is 100.

Outliers and Influential Points

In some scatter plots, there may be points that stick out. How they stick out is important in the bivariate case. **Outliers** are points that stick out from the rest of the group in a single variable. We can identify outliers in univariate data using the fence rules.

In addition to outliers, a sample may contain one or more points that are called **influential points**. Influential points are observed data points that do not follow the trend of the rest of the data. These points could have a big effect on the slope of the regression line calculation. To begin to identify an influential point, you can remove it from the dataset and see if the slope of the regression line changes significantly.

If a point is an outlier, does that necessarily make it an influential point (or vice versa)? The left graph of Figure 3.19 shows two points that stick out. One point is likely an outlier in y, but it still fits the trend. Therefore, it is not an influential point. The second point is influential (does not fit the trend) but does not appear to be an outlier in x or y. The right graph of Figure 3.19 shows a point that is both and outlier and an influential point. In summary, outliers are a univariate idea, influential points are bivariate ideas, and one does not imply the other.

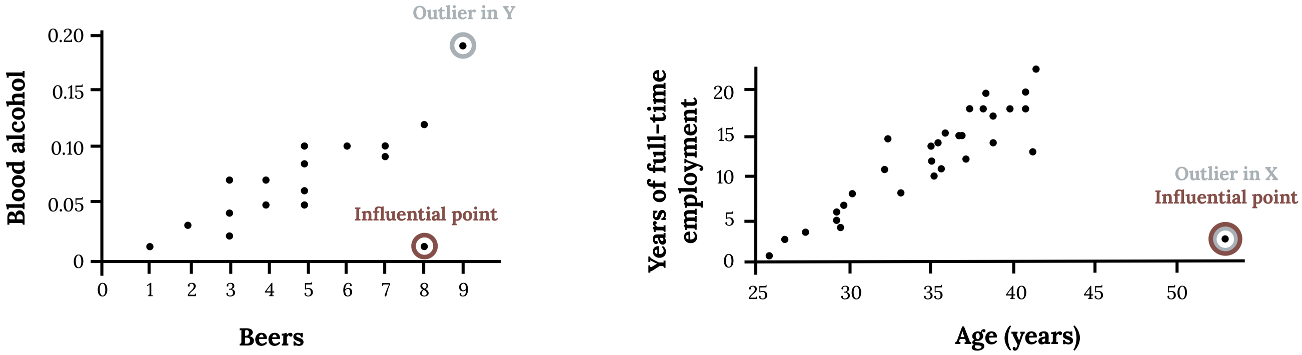


Figure 3.19: Outliers and influential points. Figure description available at the end of the section.

How do we handle these unusual points? Sometimes, they should not be included in the analysis of the data. It is possible that an outlier or influential point is a result of erroneous data. Other times, they may hold valuable information about the population under study and should remain included in the data—especially when an outlier that is not an influential point may help us to not have to extrapolate. The key is to examine carefully what causes a data point to be an outlier and/or influential point.

Identifying Outliers and/or Influential Points

Many computers and calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

We know how to find outliers in a single variable using fence rules and box plots. However, we would like some guidelines as to how far away a point needs to be in order to be considered an influential point. Outliers also have large "errors," where the error, also known as the residual, is the vertical distance from the line to the point. As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best-fit line as an outlier. The standard deviation used is the regression standard error of the residuals or errors. Another method used can be to "standardize" (more on this in CH4) the residuals, essentially calculating a Z-score. We already know that Z-scores outside of ±2 are "unusual" and we can say the same about standardized residuals.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. We can also do this numerically by calculating each residual and comparing it to twice the standard deviation. The graphical procedure is shown in the example below, followed by the numerical calculations in the next example. You would generally need to use only one of these methods.

Example

Continuing with the example from the previous section, you can determine whether there are outliers in the student exam scores If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the SSE should be smaller and the correlation coefficient ought to be closer to 1 or −1.

Here it is easy to identify the outliers graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to two or more standard deviations (±2s), then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3:

- $\hat{y} = -173.5 + 4.83x$ is the line of best fit.

- Let Y2 = −173.5 + 4.83*x* −2(16.4)
- Let Y3 = −173.5 + 4.83*x* + 2(16.4)

Notice Y2 and Y3 have the same slope as the line of best fit.

If we graph the scatter plot with the best-fit line in equation and two dotted lines (Y2 and Y3) representing ±2s from the line, you will find the point x = 65, y = 175, which is the only data point that is not between lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than two standard deviations away from the best-fit line.

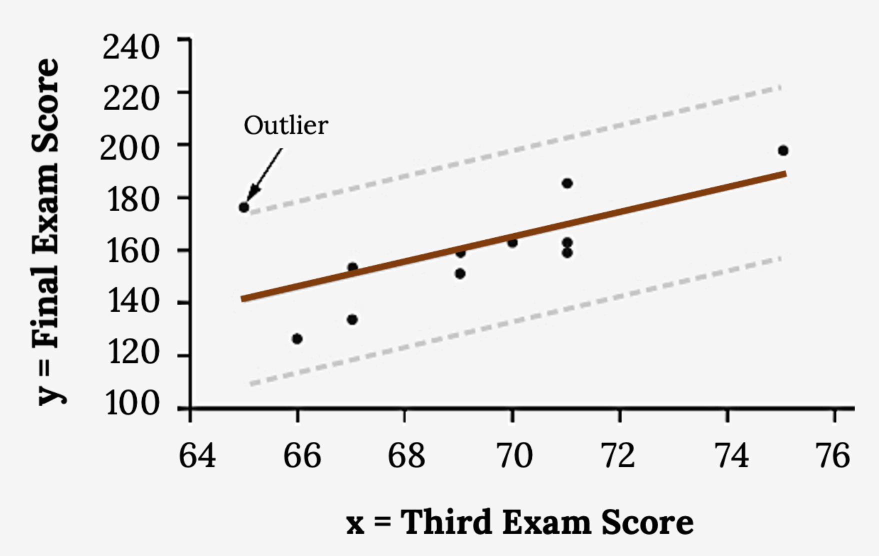


Figure 3.20: One method of identifying outliers in scatter plots. Figure description available at the end of the section.

Your Turn!

Identify the potential outlier in the scatter plot by drawing two separate lines. Suppose the standard deviation of the residuals or errors (s) is approximately s = 8.6.

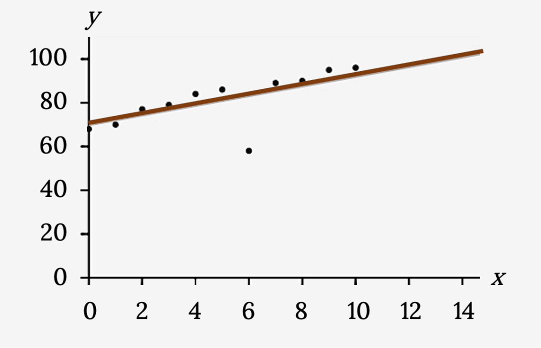


Figure 3.21: Identify the outlier. [*Figure description available at the end of the section*](#).

Residuals

In the process of numerically identifying outliers and influential points, **residuals** are one of the most important tools and are found with the formula $y_0 - \hat{y}_0 = \varepsilon_0$ (ε = the Greek letter epsilon). Though the residual is often called the error, it is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram below, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line, and the residual is positive.

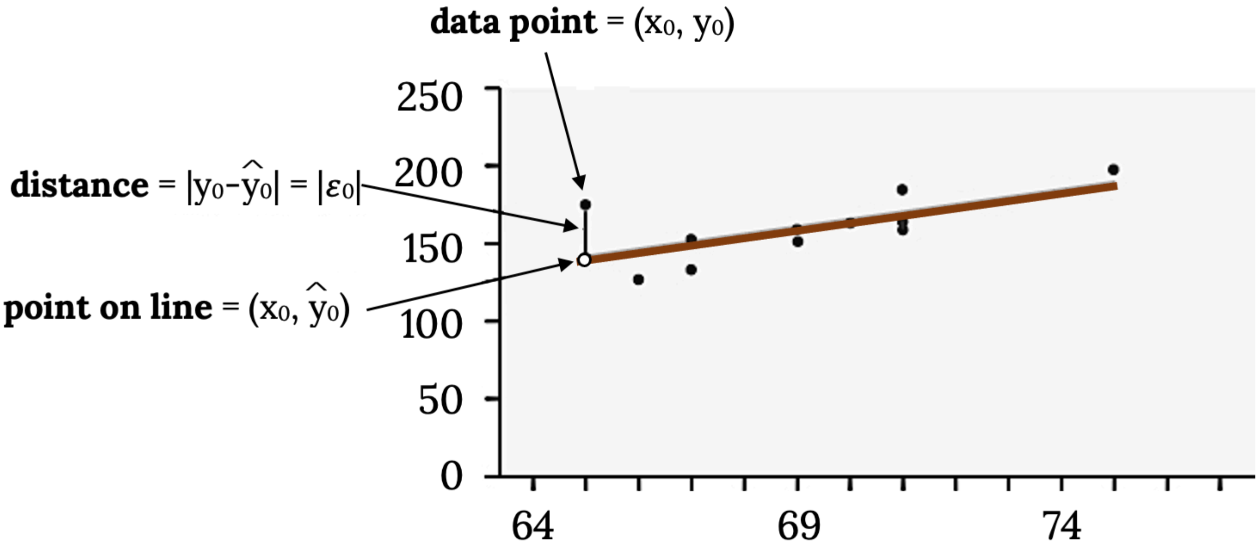


Figure 3.22: Residuals diagram. Figure description available at the end of the section.

Points that fall far from the line are points of high leverage; these points can strongly influence the slope of the least-squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line, then we consider it an influential point. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least-squares line. Let's see how to do this mathematically.

Example

For each data point, you can calculate the residuals or errors as $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, ..., 11$. Each $|\varepsilon|$ is a vertical distance. In the following table, the first two columns are the third exam and final exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y} = -173.5 + 4.83x$. The residuals have been calculated in the fourth column of the table using this formula: observed y value − predicted y value = $y - \hat{y}$.

x	*y*	*ŷ*	*y* - *ŷ*
65	175	140	175 – 140 = 35
67	133	150	133 – 150= –17
71	185	169	185 – 169 = 16
71	163	169	163 – 169 = –6
66	126	145	126 – 145 = –19
75	198	189	198 – 189 = 9
67	153	150	153 – 150 = 3
70	163	164	163 – 164 = –1
71	159	169	159 – 169 = –10
69	151	160	151 – 160 = –9
69	159	160	159 – 160 = –1

Figure 3.23: Calculating residuals

For this example, there are 11 *ε* values. If you square each ε and add, you get:

$$(\epsilon_1)^2 + (\epsilon_2)^2 + \ldots + (\epsilon_{11})^2 = \sum_{i\,=\,1}^{11} \epsilon^2$$

This is called the sum of squared errors (SSE).

For our example, the calculation is as follows:

1. First, square each |*y* – *ŷ*|.

The squares are 35^2, 17^2, 16^2, 6^2, 19^2, 9^2, 3^2, 1^2, 10^2, 9^2, and 1^2.

2. Then, add (sum) all the |y – ŷ| squared terms. Recall that $y_i - \hat{y}_i = \varepsilon_i$.

$$\sum_{i=1}^{11} \left(|y_i - y_i\,|\right)^2 = \sum_{i=1}^{11} \epsilon_i{}^2$$

= 352 + 172 + 162 + 62 + 192 + 92 + 32 + 12 + 102 + 92 + 12

= 2440 = SSE

The result, SSE, is the sum of squared errors.

s is the standard deviation of all the *y* – *ŷ* = ε values where *n* = the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{SSE}{n-2}}$$

NOTE: We divide by (*n* – 2) as the degrees of freedom (df) because the regression model involves two estimates.

For our example:

$$s = \sqrt{\frac{2440}{11-2}} = 16.47$$

NOTE: Rather than calculate these ourselves, we can find s using the computer or calculator.

More on Influential Points

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least 2s, then we would consider the data point to be “too far” from the line of best fit. We call that point a potential influential point.

Back to our example, multiply s by two:

$$(2)(16.47) = 32.94$$

This reveals that 32.94 is two standard deviations away from the mean of the $y - \hat{y}$ values.

So for this example, if any of the $|y - \hat{y}|$ values are *at least* 32.94, the corresponding (x, y) data point is a potential outlier.

We are looking for all data points for which the residual is greater than 2s = 2(16.4) = 32.8 or less than −32.8. Compare these values to the residuals in column four of the table. It appears all the $|y - \hat{y}|$'s are less than 31.29 except for the first one which is 35.

The formula $|y - \hat{y}| \geq (2)(s)$ gives us 35 > 31.29.

The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How does the outlier affect the best-fit line? Numerically and graphically, we have identified the point (65, 175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error, we should fix the error if possible or delete the data. If the data is correct, we would leave it in the dataset. For this problem, we will suppose that we examined the data and found that this outlier data was an error. As a learning experience, we will continue on and delete the outlier so that we can explore how it affects the results.

The next step is to compute a new best-fit line using the ten remaining points. The new line of best fit is $\hat{y} = -355.19 + 7.39x$, and the correlation coefficient (r) is 0.9121.

The new line with $r = 0.9121$ is a stronger correlation than the original ($r = 0.6631$) because $r = 0.9121$ is closer to one. This means that the new line is a better fit to the ten remaining data values. The line can better predict the final exam score given the third exam score. The point we deleted appears to be an influential point

It is often tempting to remove outliers and influential points. Don’t do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings—the “outliers”—they would soon go bankrupt from poorly thought-out investments.

When outliers are deleted, the researcher should either record that data was deleted and why, or the researcher should provide results both with and without the deleted data. If data is erroneous and the correct values are known (e.g., student one actually scored a 70 instead of a 65), then this correction can be made to the data.

With this new line of best fit in mind, let's revisit the remaining ten data points from the exam score example in previous sections. What would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

- Using the new line of best fit, $\hat{y}$ = −355.19 + 7.39(73) = 184.28, a student who scored 73 points on the third exam would expect to earn 184 points on the final exam.
- The original line predicted $\hat{y}$ = −173.51 + 4.83(73) = 179.08, so the prediction using the new line with the outlier eliminated differs from the original prediction.

Your Turn!

The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, *x* is the year and *y* is the CPI.

x	*y*	*x*	*y*
1915	10.1	1969	36.7
1926	17.7	1975	49.3
1935	13.7	1979	72.6
1940	14.7	1980	82.4
1947	24.1	1986	109.6
1952	26.5	1991	130.7
1964	31.0	1999	166.6

Figure 3.24: CPI Values

a. Draw a scatter plot of the data.
b. Calculate the least-squares line. Write the equation in the form $\hat{y} = a + bx$.
c. Draw the line on the scatter plot.
d. Find the correlation coefficient.

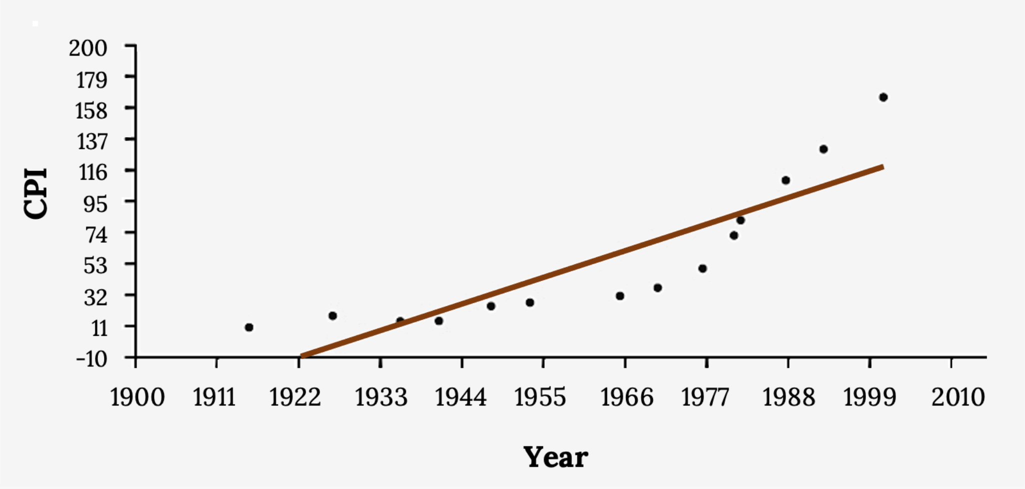


Figure 3.25: Scatter plot of CPI values. *[Figure description available at the end of the section](#)*.

e. What is the average CPI for the year 1990?
f. Comment on the appropriateness of this linear model. Do there appear to be any outliers or influential points?

Solution

a. The scatter plot should look similar to the one in (d).

b. $\hat{y}$ = −3204 + 1.662x is the equation of the line of best fit.

c. [See image]

d. r = 0.8694

e. $\hat{y}$ = −3204 + 1.662(1990) = 103.4 CPI

f. There are no outliers or influential points in the example. Notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatterplot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician should use other methods to fit a curve to this data, rather than model the data with a line. In addition to doing the calculations, it is always important to look at the scatterplot when deciding whether a linear model is appropriate.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 3.19: Kindred Grey (2024). *Outliers and influential points*. CC BY-SA 4.0.

Figure 3.20: Kindred Grey (2020). *One method of identifying outliers in scatter plots*. CC BY-SA 4.0. Adaptation of Figure 12.18 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-6-outliers

Figure 3.21: Kindred Grey (2020). *Identify the outlier.* CC BY-SA 4.0. Adaptation of Figure 12.19 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-6-outliers

Figure 3.22: Kindred Grey (2020). *Residuals diagram.* CC BY-SA 4.0. Adaptation of Figure 12.10 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation

Figure 3.25: Kindred Grey (2020). *Scatter plot of CPI values*. CC BY-SA 4.0. Adaptation of Figure 12.20 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/12-6-outliers

Figure Descriptions

Figure 3.19: Left: scatterplot with an outlier in Y (there's a point way high up above the rest) and an influential point (one that is not in line with the other points). Right: scatterplot where one point is both an outlier in x and an influential point

Figure 3.20: The same scatter plot of exam scores with a line of best fit. Two dashed lines run parallel to the line of best fit. The dashed lines run above and below the best fit line at equal distances. One data point falls outside the boundary created by the dashed lines—it is an outlier.

Figure 3.21: Scatterplot with dots in an almost perfect line from bottom left corner to top right corner of graph. There is one dot that does not follow this linear pattern.

Figure 3.22: The same scatter plot of exam scores with a line of best fit. One data point is highlighted along with the corresponding point on the line of best fit directly beneath it. Both points have the same x-coordinate. The distance between these two points illustrates how to compute the sum of squared errors.

Figure 3.25: Scatter plot and line of best fit of the consumer price index data, on the y-axis, and year data, on the x-axis. Moderately strong positive linear correlation.

Chapter 3 Wrap-Up

Concept Check

Take this quiz to check your comprehension of this chapter.

If you are using an offline version of this text, access the quiz for this chapter via the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=406#h5p-185

Section Resources

If you are using an offline version of this text, access these materials by visiting https://doi.org/10.7294/26207456.

3.1 Introduction to Bivariate Data

3.2 Visualizing Bivariate Quantitative Data

3.3 Measures of Association

3.4 Modeling Linear Relationships

3.5 Cautions about Regression

Key Terms

Try to define the terms below on your own. Check your response by clicking on the term, or looking at the end-of-book glossary!

3.1 Introduction to Bivariate Data

- **Contingency table**

3.2 Visualizing Bivariate Quantitative Data

- **Bivariate data**
- **Response variable**
- **Explanatory variable**

3.3 Measures of Association

- **Correlation coefficient (r)**
- **Coefficient of determination (r^2)**

3.4 Modeling Linear Relationships

- **Linear regression**
- **Slope**
- **y-intercept**

3.5 Cautions about Regression

- **Extrapolation**
- **Outliers**
- **Influential points**
- **Residuals**

Extra Practice

Extra practice problems are available at the end of the book (Chapter 3 Extra Practice).

CHAPTER 4: PROBABILITY DISTRIBUTIONS

4.1 Introduction to Probability and Random Variables

Learning Objectives

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

- Understand the terminology and basic rules of probability
- Handle general discrete random variables
- Recognize and apply the binomial distribution
- Understand general continuous random variables
- Recognize and apply special cases of continuous random variables (uniform, normal)
- Use the normal distribution to approximate the binomial

More than likely, you have used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Figure 4.1: Meteor shower. Meteor showers are rare, but the probability of them occurring can be calculated. [Figure description available at the end of the section](#).

Probability

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An experiment is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **probability experiment**. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are listing the possible outcomes, creating a tree diagram, or creating a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, S = {*H*, *T*} where *H* (heads) and *T* (tails) are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written P(A).

The probability of any outcome is the long-term relative frequency of that outcome. Probabilities are between zero and one, inclusive (that is, zero, one, and all numbers between these values). P(A) = 0 means that event A can never happen. P(A) = 1 means that event A always happens. P(A) = 0.5 means that event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times), the relative frequency of heads approaches 0.5 (the probability of heads).

A **probability model** is a mathematical representation of a random process that lists all possible outcomes and assigns probabilities to each of them. This type of model is our ultimately our goal when moving forward in our study of statistics.

The Law of Large Numbers

An important characteristic of probability experiments known as the **law of large numbers** states that, as the number of repetitions of an experiment increases, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word "empirical" is often used instead of the word "observed.")

If you toss a coin and record the result, what is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. Probability does not describe the short-term results of an experiment; rather, it gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the law of large numbers.

The Axioms of Probability

Finding probabilities in more complicated situations starts with the three axioms of probability:

1. $P(S) = 1$
2. $0 \leq P(E) \leq 1$
3. For each two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

The first two axioms should be fairly intuitive. Axiom 1 says that the probabilities of all outcomes in a sample space will always add up to 1. Axiom 2 says the probability of any event must be between 0 and 1. For now, the third axiom, called the disjoint addition rule, isn't that important, but the upcoming ideas are based on the first two axioms.

The Complement

Suppose we know the probability of an event occurring but want to know the probability it does not occur, or vice versa? We can easily find this from the first two axioms of probability.

We call all of the outcomes in a sample space that are NOT included in an event the **complement** of the event. The complement of event A is usually denoted by $\overline{A}$, A′ (read "A prime"), or A^C.

There are several useful forms of the complement rule:

- P(A) + P(A′) = 1
- 1 – P(A) = P(A′)
- 1 – P(A') = P(A)

Example

If S = {1, 2, 3, 4, 5, 6} and A = {1, 2, 3, 4}, then A′ = {5, 6}:

P(A) = $\frac{4}{6}$, P(A′) = $\frac{2}{6}$, and P(A) + P(A′) = $\frac{4}{6} + \frac{2}{6}$ = 1

Random Variables

Random variables (RVs) are probability models quantifying situations. A random variable describes the outcomes of a statistical experiment in words or as a function that assigns each element of a sample space a unique real number. Uppercase letters such as X or Y typically denote a random variable. Lowercase letters like x or y denote a specific value of that random variable. If X is a random variable, then X is written in words, and x is given as a number. For example, the probability of the random variable X being equal to 3 is denoted as P(X=3).

There are both continuous and discrete random variables depending on the type of data that situation would produce. We will begin with **discrete random variables (DRVs)** and revisit **continuous random variables (CRVs)** in the future.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 4.1: Ed Sweeney (2009). 2009 *Leonid Meteor*. CC BY 2.0. https://flic.kr/p/7girE8

Figure Descriptions

Figure 4.1: Photo of the night sky. A meteor and its tail are shown entering the earth's atmosphere.

4.2 Discrete Random Variables

A student takes a ten-question, true-false quiz. Because of the student's busy schedule, they could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during peak hours?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A random variable describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Discrete Random Variables

We have previously seen the word discrete associated with types of data. Discrete means we have a countable number of outcomes, so a **discrete random variable** is an RV that models a process or experiment that produces discrete data.

For instance, let X stand for the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT*, *THH*, *HTH*, *HHT*, *HTT*, *THT*, *TTH*, *HHH*. Then, x = 0, 1, 2, 3. X is given in words, while x is given in numbers. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, and 3), X is a discrete random variable.

Example

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X represent the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, x = 0, 1, 2, 3, 4, 5.

P(x) = probability that X takes on a value x.

x	P(x)
0	$P(x=0)=\frac{2}{50}$
1	$P(x=1)=\frac{11}{50}$
2	$P(x=2)=\frac{23}{50}$
3	$P(x=3)=\frac{9}{50}$
4	$P(x=4)=\frac{4}{50}$
5	$P(x=5)=\frac{1}{50}$

Figure 4.2: Newborn baby crying

Is this a valid discrete probability distribution?

Solution

X takes on the values 0, 1, 2, 3, 4, and 5. This ***is*** a valid discrete PDF because:

Each P(x) is between zero and one, inclusive.

The sum of the probabilities is one, that is,

$\frac{2}{50}+\frac{11}{50}+\frac{23}{50}+\frac{9}{50}+\frac{4}{50}+\frac{1}{50}=1$

Your Turn!

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X represent the number of times a patient rings the nurse during a 12-hour shift. For this exercise, x = 0, 1, 2, 3, 4, 5. P(x) = the probability that X takes on value x. Is this a discrete probability distribution function? If so, provide two reasons why this is or is not the case.

X	P(x)
0	$P(x=0)=\frac{4}{50}$
1	$P(x=1)=\frac{8}{50}$
2	$P(x=2)=\frac{16}{50}$
3	$P(x=3)=\frac{14}{50}$

X	P(x)
4	$P(x=4)=\frac{6}{50}$
5	$P(x=5)=\frac{2}{50}$

Figure 4.3: Post-op patients

Characteristics and Notation

The distribution of a discrete random variable is often pictured in a table, but it may also be represented by a graph or formula. There are two main characteristics it should exhibit:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

The **probability mass function (PMF)** of a DRV tells you the probability of a random variable taking on a certain value. Notation-wise, this means P(X = *x*). This is also sometimes (erroneously) called probability distribution function (PDF).

The **cumulative distribution function (CDF)** of a DRV tells you the probability of random variable being less than or equal to a certain value. Notation-wise, this means P(X ≤ *x*).

A probability distribution function is a pattern. You try to fit a probability problem into a pattern or distribution in order to perform the necessary calculations. These distributions are tools that make it easier to solve probability problems. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

Example

Suppose Nancy has classes three days a week. She attends all three days 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

Let X represent the number of days Nancy ______.

Solution
attends class per week

X takes on what values?

Solution
0, 1, 2, and 3

Construct a probability distribution table (called a PDF table) like the one below for the week chosen at random. The table should have two columns labeled *x* and *P*(*x*). What does the *P*(*x*) column sum to?

x	$P(x)$
0	
1	
2	
3	

Figure 4.4: Blank PDF

Solution

x	$f(x)$
0	0.01
1	0.04
2	0.15
3	0.80

Figure 4.5: Blank table

Construct the cumulative probability distribution function.

Solution

x	$f(x)$	$F(x)$
0	0.01	0.01
1	0.04	0.05
2	0.15	0.20
3	0.80	1

Figure 4.6: Cumulative PDF

Your Turn!

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X, and what values does it take on?

Measures of Discrete Random Variables

Once we know how to work with discrete random variables, we may be interested in some other measures of the data, such as the mean, variance, and standard deviation. These ideas resurface here, but in the slightly different context of random variables.

The Expected Value (Mean) of a Discrete Random Variable

Recall the law of large numbers, which states as the number of trials in a probability experiment increases, our results become closer to what we expect. When evaluating the long-term results of statistical experiments, we often want to know the “average” outcome. This long-term average is known as the mean or **expected value** of the random variable and is denoted by the Greek letter μ or E[X] in the context of random variables. In other words, this is the average value you would expect after conducting many trials of an experiment.

To find the expected value or long-term average, we simply multiply each value of the random variable by its probability and add the products. It is essentially a probability weighted average of the values of the random variable.

Mean or expected value:

$$\mu = \sum_{x\in X} xP\left(x\right)$$

Example

A men’s soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ, of the number of days per week the men’s soccer team plays soccer.

To do the problem, first let the random variable X represent the number of days the men’s soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table, adding a column x*P(x). In this column, you will multiply each *x* value by its probability. This table is called an expected value table. The table helps you calculate the expected value, or long-term average.

x	P(x)	x*P(x)
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Figure 4.7: Expected value table

What is the expected value?

Solution

Add the last column *x***P*(*x*) to find the long term average or expected value.

(0)(0.2) + (1)(0.5) + (2)(0.3) = 0 + 0.5 + 0.6 = 1.1

The expected value is 1.1. The men's soccer team would, on average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week. We say μ = 1.1.

Your Turn!

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

x	P(x)
0	$P(x=0)=\frac{4}{50}$
1	$P(x=1)=\frac{8}{50}$
2	$P(x=2)=\frac{16}{50}$
3	$P(x=3)=\frac{14}{50}$
4	$P(x=4)=\frac{6}{50}$
5	$P(x=5)=\frac{2}{50}$

Figure 4.8: Post-op patients

The Variance and Standard Deviation of a Discrete Random Variable

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root.

Finding the variance (σ^2 or V[X]) and standard deviation (σ or SD[X]) of a random variable starts similarly to finding these measures for a data sample, which we have seen before. However, the process differs at its fourth step and looks more like a probability weighted average of the squared deviations similar to the method used to calculate an expected value:

1. Find the mean.
2. Subtract the mean from each value of *x* to get your deviations.
3. Square each deviation.
4. Multiply each squared deviation by its probability, *P*(*x*).
5. Sum each of the products.

At this point, we now have the variance and can take the square root of the variance to get our standard deviation. The formula looks like this:

$$\sigma = \sqrt{\sum_{x \in X} (x-\mu)^2 P(x)}$$

Example

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

x	P(x)	x*P(x)	(x – μ)^2 · P(x)
0	$P(x=0) = \frac{2}{50}$	$(0)(\frac{2}{50}) = 0$	$(0-2.1)^2 \cdot 0.04 = 0.1764$
1	$P(x=1) = \frac{11}{50}$	$(1)(\frac{11}{50}) = \frac{11}{50}$	$(1-2.1)^2 \cdot 0.22 = 0.2662$
2	$P(x=2) = \frac{23}{50}$	$(2)(\frac{23}{50}) = \frac{46}{50}$	$(2-2.1)^2 \cdot 0.46 = 0.0046$
3	$P(x=3) = \frac{9}{50}$	$(3)(\frac{9}{50}) = \frac{27}{50}$	$(3-2.1)^2 \cdot 0.18 = 0.1458$
4	$P(x=4) = \frac{4}{50}$	$(4)(\frac{4}{50}) = \frac{16}{50}$	$(4-2.1)^2 \cdot 0.08 = 0.2888$
5	$P(x=5) = \frac{1}{50}$	$(5)(\frac{1}{50}) = \frac{5}{50}$	$(5-2.1)^2 \cdot 0.02 = 0.1682$

Figure 4.9: Newborn baby crying

You expect a newborn to wake its mother after midnight 2.1 times per week on the average.

Add the values in the third column of the table to find the expected value of X.

Solution
μ = expected value = $\frac{105}{50}$ = 2.1

Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value *x*, multiply the square of its deviation by its probability. Each deviation has the format *x* – μ.

Add the values in the fourth column of the table.

Solution
0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05

The standard deviation of X is the square root of this sum.

Solution
σ = $\sqrt{1.05}$ ≈ 1.0247

The mean, μ, of a discrete probability function is the expected value.

Solution
μ = $\Sigma(x \cdot P(x))$

The standard deviation, σ, of the PDF is the square root of the variance.

Solution
σ = $\sqrt{\Sigma\left[(x-\mu)^2 \cdot \mathrm{P}(x)\right]}$

When all outcomes in the probability distribution are equally likely, these formulas coincide with the mean and standard deviation of the set of possible outcomes.

Your Turn!

On May 11, 2013, at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. You bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win $100. If you lose the bet, you pay $10. If X is the amount of profit from a bet, find the mean and standard deviation of X.

Note on Calculations

For probability distributions, we generally use a calculator or a computer to calculate μ and σ to reduce roundoff error. For many special cases of probability distributions, there are shortcut formulas for calculating μ, σ, and associated probabilities. We will see some of these in the future.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

4.3 The Binomial Distribution

We have seen how to deal with general discrete random variables, but there are also special cases of DRVs. If we can identify them, they can provide us some insight and shortcuts. The first of these is the **binomial distribution**.

The Binomial Setting

There are three characteristics of a **binomial experiment**:

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes for each trial: success and failure. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. Note that $p + q = 1$.
3. The n trials are **independent** and are repeated using identical conditions. Because they are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that, for each individual trial, the probability of a success (p) and probability of a failure (q) remain the same.

Let's say that the withdrawal rate from an elementary physics course at ABC College is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. In this instance, the random variable X represents the number of students who withdraw from the randomly selected elementary physics class.

Any experiment that has the second and third characteristics listed above and where $n = 1$ is called a **Bernoulli trial** (named after Jacob Bernoulli who studied them extensively in the late 1600s). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli trials.

For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. If Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$, then $q = 0.4$. This means that, for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same. This situation meets the binomial requirements.

In contrast, the following example illustrates a problem that is not binomial, as it violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn without replacement. The first name drawn determines the chairperson, and the second name, the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$ when the first draw selects a student. The probability is $\frac{6}{15}$ when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

Example

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

This is a binomial problem because there is only a success or a ____________, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.

Solution
failure

If we are interested in the number of students who do their homework on time, then how do we define X?

Solution
X = the number of statistics students who do their homework on time

What values does *x* take on?

Solution
0, 1, 2, …, 50

What is a "failure," in words?

Solution
Failure is defined as a student who does not complete their homework on time.

The probability of a success is $p = 0.70$. The number of trials is $n = 50$.

If $p + q = 1$, then what is q?

Solution
$q = 0.30$

The words "at least" translate as what kind of inequality for the probability question $P(x$ __ $40)$.

Solution
Greater than or equal to (≥)

The probability question is $P(x \geq 40)$.

Your Turn!

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Can we use the binomial here?

Notation for the Binomial

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X counts the number of successes obtained in n independent trials.

$$X \sim B(n, p)$$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p (n = number of trials, p = probability of a success on each trial).

Since the binomial counts the number of successes, x, in n trials, the range of values for a binomial random variable could be anything from 0 to n (x = 0, 1, 2, …, n).

Binomial Probability Function

Once we have decided the binomial is applicable for a given situation, we can use the binomial probability function to find the probability of a specific number of successes, P(X = x). The binomial **probability mass function (PMF)** is made up of two parts.

First, we need to find out how many different ways we can get x successes in n trials. To do this, we can use the choose function, also called the binomial coefficient, written as:

$$nC_X = C_x^n = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

NOTE: The ! mark is the factorial operator.

The next part gives us the probability of a single one of those ways to get x successes in n trials. We can achieve this by using our independent multiplication rule, in which we multiply the probability of success (p) raised to the number of successes (x) by the probability of failure (q = 1 − p) raised to the number of failures (n − x).

$$p^{x}q^{(n-x)}$$

Since we know each of these ways are equally likely and how many ways are possible, we can now put the two pieces together. We multiply the probability of one way by how many we have, giving us our overall probability of x successes in n trials.

$$P(X = x) = \frac{n!}{x!(n-x)!}p^{x}q^{(n-x)}$$

Unfortunately the binomial does not have a nice form of **cumulative distribution function (CDF)**, but it is simply the sum of PDFs up until that point. Consider the following example to demonstrate this point.

Example

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. Twenty adult workers are randomly selected.

Let X represent the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, …, 20, where n = 20, p = 0.41, and q = 1 – 0.41 = 0.59.

X ~ B(20, 0.41)

The y-axis contains the probability of x, where X is the number of workers who have only a high school diploma.

The graph of X ~ B(20, 0.41) is as follows:

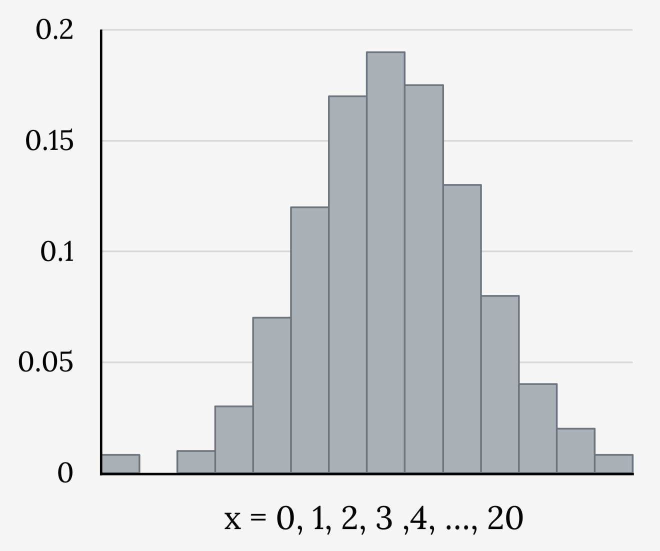


Figure 4.10: Workers with diplomas. Figure description available at the end of the section.

Find the probability that exactly 12 of them have a high school diploma.

Solution
We can simply plug into the binomial PMF

P(X = x) = $\frac{n!}{x!(n-x)!}p^{x}q^{(n-x)}$

for P(X=12) with n=20 and p=0.41, $\frac{20!}{12!(20-12)!}0.41^{12}(1\text{-}0.41)^{(20-12)}$

Find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma without pursuing any further education?

Solution
If you want to find P(x = 12), use the pdf (binompdf). If you want to find P(x > 12), use 1 – binomcdf(20, 0.41, 12).

The probability that at most 12 workers have a high school diploma but do not pursue any further education is 0.9738.

Your Turn!

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find:

(a) The probability that exactly 14 of them participate in a community volunteer program outside of school. First, try plugging in to the binomial formula by hand, then check yourself with technology.

(b) The probability that exactly 14 of them participate in a community volunteer program outside of school. Rely on technology for this cumulative probability.

Measures of the Binomial Distribution

The mean, μ, and variance, σ^2, for the binomial probability distribution are μ = np and σ^2 = npq. The standard deviation, σ, is then σ = $\sqrt{npq}$.

Example

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X represent the number of pages that feature signature artists.

1. What values does *x* take on?
2. What is the probability distribution? Find the following probabilities:
    a. The probability that two pages feature signature artists
    b. The probability that at most six pages feature signature artists
    c. The probability that more than three pages feature signature artists
3. Using the formulas, calculate the mean and standard deviation.

Solution

1. x = 0, 1, 2, 3, 4, 5, 6, 7, 8

2. X ~ $B(100, \frac{8}{560})$

a. $P(x = 2)$ = binompdf(100, $\frac{8}{560}$, 2) = 0.2466

b. $P(x \leq 6)$ = binomcdf(100, $\frac{8}{560}$, 6) = 0.9994

c. $P(x > 3) = 1 - P(x \leq 3)$ = 1 − binomcdf(100, $\frac{8}{560}$, 3) = 1 − 0.9443 = 0.0557

3. Mean = $np = (100)(\frac{8}{560}) = \frac{8}{560} \approx 1.4286$

Standard deviation = $\sqrt{npq} = \sqrt{(100)(\frac{8}{560})(\frac{552}{560})} \approx 1.1867$

Your Turn!

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X represent the number of American adults out of a random sample of 50 who prefer saving to spending.

1. What is the probability distribution for X?
2. Use your calculator to find the following probabilities:
    a. The probability that 25 adults in the sample prefer saving over spending
    b. The probability that at most 20 adults prefer saving
    c. The probability that more than 30 adults prefer saving
3. Using the formulas, calculate the mean and standard deviation of X.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 4.10: Kindred Grey (2020). *Workers with diplomas.* CC BY-SA 4.0.

Figure Descriptions

Figure 4.10: Histogram showing a binomial probability distribution. It is made up of bars that are fairly normally distributed. The x-axis shows values from zero to 20. The y-axis shows values from zero to 0.2 in increments of 0.05.

4.4 Continuous Random Variables

Continuous random variables (CRVs) have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the lifespan of a computer chip, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

NOTE: The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X, which is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by the area under the curve.

The curve—represented by the symbol *f*(*x*)—is called the **probability density function (PDF)**. *f*(*x*) is the function that corresponds to the graph; we use the density function *f*(*x*) to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function (CDF)**. The cumulative distribution function is used to evaluate probability as area.

When dealing with CDFs:

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c < x < d)$ is the area under the curve, above the x-axis, to the right of c, and the left of d.
- The probability that x takes on any single individual value is zero: $P(x = c) = 0$. The area below the curve, above the x-axis, and between $x = c$ and $x = c$ has no width and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$ is the same as $P(c \le x \le d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions; however, much of the work has already been done for us. The formulas to find the area in this textbook have already been found by using the techniques of integral calculus.

Some Continuous Distributions

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way. We do not often handle general CRVs, instead studying special known cases most of the time. The following graphs illustrate some of these these distributions.

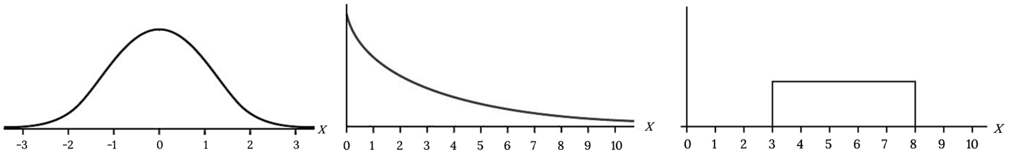


Figure 4.11: Continuous distributions. Figure description available at the end of the section.

Probability Density Functions

We begin by defining a continuous probability density function, using the function notation $f(x)$. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. For continuous probability distributions, you can think about it as: PROBABILITY = AREA.

The Uniform Distribution

The (continuous) **uniform distribution** is fairly simple and is a great place to start in demonstrating the ideas of continuous distributions. It is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data is inclusive or exclusive of endpoints.

The notation for the uniform distribution is:

X ~ U(a, b), where a = the lowest value of x and b = the highest value of x.

The probability density function is:

$$f(x) = \frac{1}{b-a} \text{ for } a \le x \le b.$$

Formulas for the theoretical mean and standard deviation are:

$$\mu = \frac{a+b}{2} \text{ and } \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Example

Suppose a researcher tracked the duration each time his eight-week-old baby smiled. The following data are 55 smiling times, in seconds.

Smiling times (in seconds)										
10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

Figure 4.12: Smiling times

The sample mean is 11.49, and the sample standard deviation is 6.23.

We will assume that the smiling times, in seconds, follow a uniform distribution between zero and 23 seconds, inclusive. This means that any smiling time from zero to and including 23 seconds is equally likely. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let X represent length, in seconds, of an eight-week-old baby's smile.

For this example:

$$X \sim U(0, 23) \text{ and } f(x) = \frac{1}{23-0} \text{ for } 0 \le X \le 23.$$

For this problem, the theoretical mean and standard deviation are:

$$\mu = \frac{0 + 23}{2} = 11.50 \text{ seconds and } \sigma = \sqrt{\frac{(23 - 0)^2}{12}} = 6.64 \text{ seconds.}$$

Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation in this example.

Example

Consider the function $f(x) = \frac{1}{20}$ for $0 \le x \le 20$.

- x = a real number
- The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \le x \le 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.

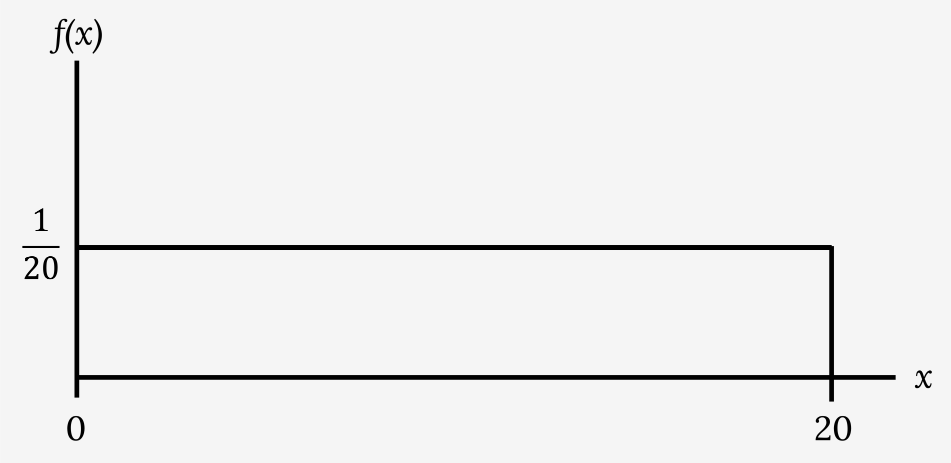


Figure 4.13: Example of a function. Figure description available at the end of the section.

- $f(x) = \frac{1}{20}$ for $0 \le x \le 20$.
- The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \le x \le 20$.
- The area between $f(x) = \frac{1}{20}$ where $0 \le x \le 20$ and the x-axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.
- Area = $20(\frac{1}{20}) = 1$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $0 < x < 2$.

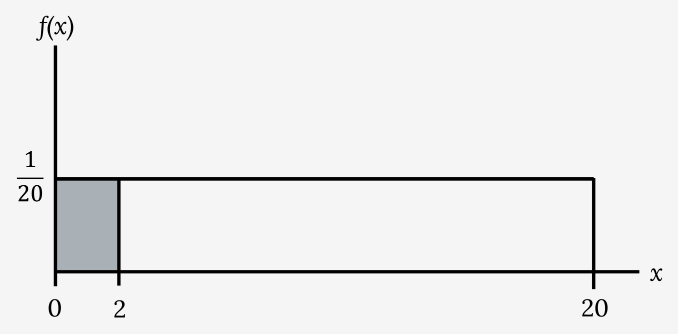


Figure 4.14: Finding area. Figure description available at the end of the section.

Solution

Area = $(2 - 0)(\frac{1}{20}) = 0.1$

$(2-0) = 2$ = base of a rectangle

Reminder: area of a rectangle = (base)(height). The area corresponds to a probability.

The probability that x is between zero and two is 0.1, which can be written mathematically as $P(0 < x < 2) = P(x < 2) = 0.1$.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $4 < x < 15$.

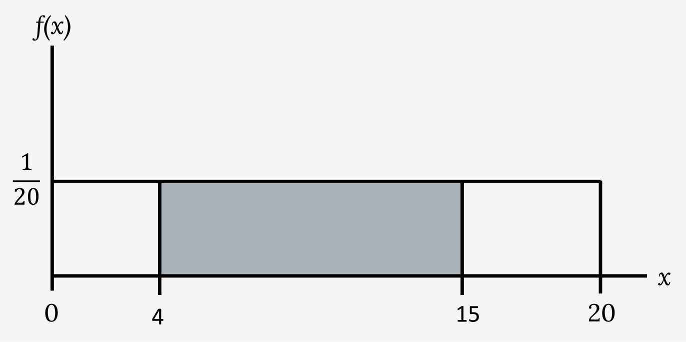


Figure 4.15: Finding area. Figure description available at the end of the section.

Solution

Area = $(15 - 4)(\frac{1}{20}) = 0.55$

$(15 - 4) = 11$ = base of a rectangle

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

Suppose we want to find P(x = 15).

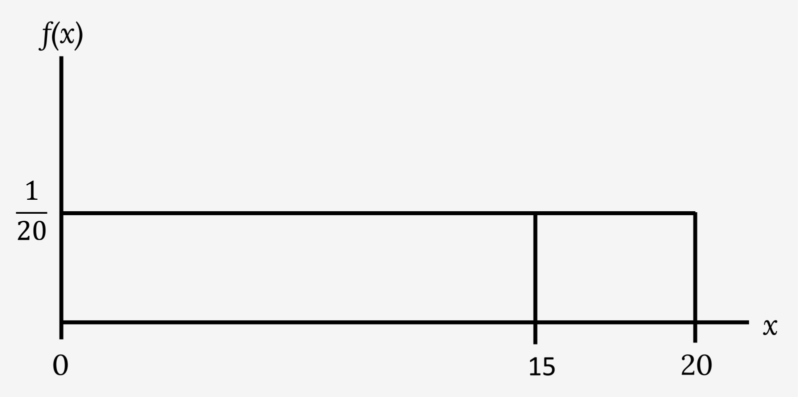


Figure 4.16: Finding a value. Figure description available at the end of the section.

Solution

On an x-y graph, x = 15 is a vertical line. A vertical line has no width (or zero width). Therefore, P(x = 15) = (base)(height) = (0)($\frac{1}{20}$) = 0.

Recall that P(X ≤ x), which can also be written as P(X < x) for continuous distributions, is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can also use the CDF to calculate P(X > x). The CDF gives "area to the left," and P(X > x) gives "area to the right." We calculate P(X > x) for continuous distributions as follows: P(X > x) = 1 – P (X < x).

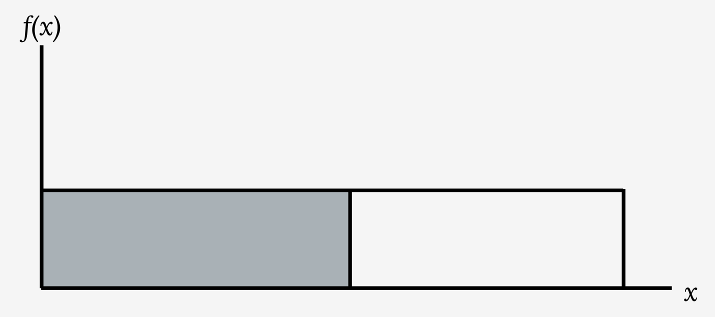


Figure 4.17: Area to the left. Figure description available at the end of the section.

Label the graph with $f(x)$ and x. Scale the x- and y-axes with the maximum x and y values.

Solution

$f(x) = \frac{1}{20}$, $0 \le x \le 20$.

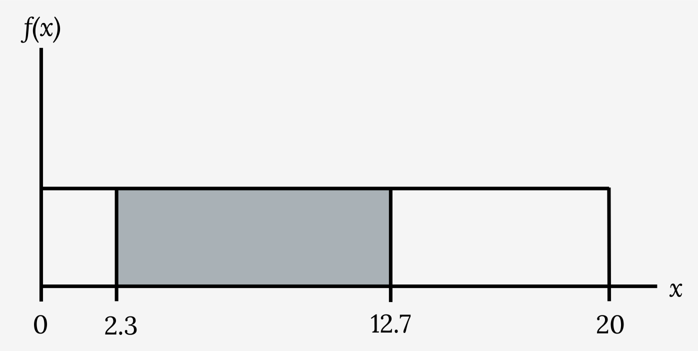


Figure 4.18: Finding area. [Figure description available at the end of the section](#).

To calculate the probability that x is between two values, look at the graph above. Shade the region between x = 2.3 and x = 12.7. Then calculate the shaded area of a rectangle.

Solution

P(2.3<x<12.7) = (base)(height) = (12.7−2.3)($\frac{1}{20}$) = 0.52

Your Turn!

Consider the function $f(x) = \frac{1}{8}$ for $0 \leq x \leq 8$. Draw the graph of $f(x)$, and find $P(2.5 < x < 7.5)$.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 4.11: Kindred Grey (2020). *Continuous distributions*. CC BY-SA 4.0. Adaptation of Figures 5.37, 5.38,

and 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/5-practice

Figure 4.13: Kindred Grey (2020). *Example of a function.* CC BY-SA 4.0.

Figure 4.14: Kindred Grey (2020). *Finding area.* CC BY-SA 4.0.

Figure 4.15: Kindred Grey (2020). *Finding area.* CC BY-SA 4.0.

Figure 4.16: Kindred Grey (2020). *Finding a value.* CC BY-SA 4.0.

Figure 4.17: Kindred Grey (2020). *Area to the left.* CC BY-SA 4.0.

Figure 4.18: Kindred Grey (2020). *Finding area.* CC BY-SA 4.0.

Figure Descriptions

Figure 4.11: Three distributions side by side. Left: Bell-shaped graph; the symmetric graph reaches maximum height at x = zero and slopes downward gradually to the x-axis on each side of the peak. Middle: downward sloping line graph; it begins at a point on the y-axis and approaches the x-axis at the right edge of the graph. Right: The horizontal axis ranges from zero to 10. The distribution is modeled by a rectangle extending from x = three to x = eight.

Figure 4.13: This shows the graph of the function f(x) = 1/20. A horizontal line ranges from the point (0, 1/20) to the point (20, 1/20). A vertical line extends from the x-axis to the end of the line at point (20, 1/20) creating a rectangle.

Figure 4.14: This shows the graph of the function f(x) = 1/20. A horizontal line ranges from the point (0, 1/20) to the point (20, 1/20). A vertical line extends from the x-axis to the end of the line at point (20, 1/20) creating a rectangle. A region is shaded inside the rectangle from x = zero to x = two.

Figure 4.15: This shows the graph of the function f(x) = 1/20. A horizontal line ranges from the point (0, 1/20) to the point (20, 1/20). A vertical line extends from the x-axis to the end of the line at point (20, 1/20) creating a rectangle. A region is shaded inside the rectangle from x = four to x = 15.

Figure 4.16: This shows the graph of the function f(x) = 1/20. A horizontal line ranges from the point (0, 1/20) to the point (20, 1/20). A vertical line extends from the x-axis to the end of the line at point (20, 1/20) creating a rectangle. A vertical line extends from the horizontal axis to the graph at x = 15.

Figure 4.17: This shows the graph of the function f(x) = 1/20. A horizontal line ranges from the point (0, 1/20) to the point (20, 1/20). A vertical line extends from the x-axis to the end of the line at point (20, 1/20) creating a rectangle. The area to the left of a value, x, is shaded.

Figure 4.18: This shows the graph of the function f(x) = 1/20. A horizontal line ranges from the point (0, 1/20) to the point (20, 1/20). A vertical line extends from the x-axis to the end of the line at point (20, 1/20) creating a rectangle. A region is shaded inside the rectangle from x = 2.3 to x = 12.7.

4.5 The Normal Distribution

The **normal (Gaussian) distribution** is the most important of all the distributions, continuous or otherwise. Its graph is symmetric, bell-shaped, and unimodal. It is widely used and even more widely abused. You see this distribution in almost all disciplines, including psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. In both the natural world and in human society, many elements—from IQ scores to real estate prices—fit a normal distribution.

The normal distribution has two parameters (i.e., two numerical descriptive measures): the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing X ~ N(μ, σ).

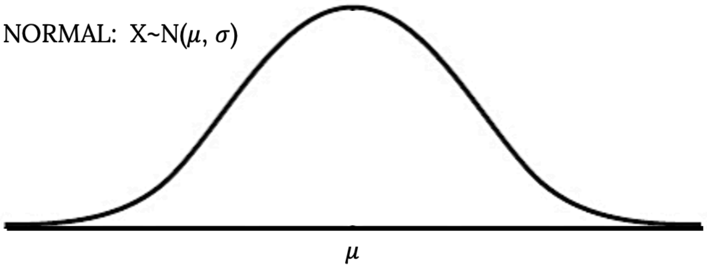


Figure 4.19: Normal distribution. Figure description available at the end of the section.

The probability density function of this curve is as follows:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \times \ e^{-\frac{1}{2} \times \left(\frac{x-\mu}{\sigma}\right)^2}$$

where:

- $-\infty < X < \infty$
- $-\infty < \mu < \infty$
- $\sigma > 0$

As you can see, the normal PDF is a rather complicated function. This could be a problem since the normal distribution is so widely used. However, we will see some ways we can work around this.

The cumulative distribution function is P(X ≤ *x*). It can be calculated either by calculus, technology, or a table (though technology has made tables almost obsolete).

The curve is symmetric about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve, which becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the standard normal distribution.

The Empirical Rule

The place to start when working with the normal distribution is the **empirical rule**. It applies to any normal distribution or data that has a bell-shaped, symmetric curve. According to the rule, if X is a random variable and has a normal distribution with mean μ and standard deviation σ, then:

- Approximately 68% of the values of x are within one standard deviation of the mean ($\pm\sigma$ *or* z-scores of ±1)
- Approximately 95% of the values of x are within two standard deviations of the mean ($\pm 2\sigma$ *or* z-scores of ±2)
- Approximately 99.7% of the values of x are within three standard deviations of the mean ($\pm 3\sigma$ *or* z-scores of ±3)

The empirical rule is also known as the 68-95-99.7 rule.

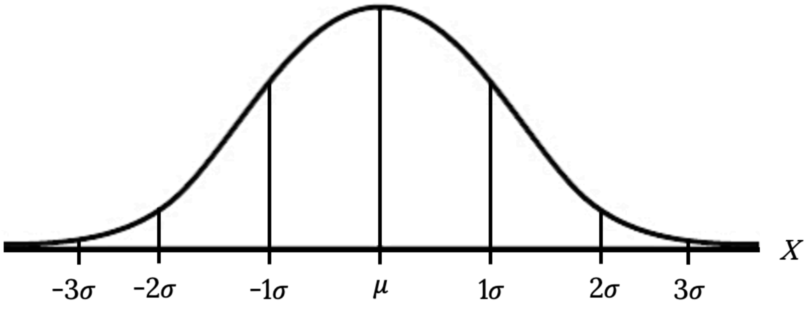


Figure 4.20: Empirical rule. Figure description available at the end of the section.

Example

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie within one standard deviation of the mean. Therefore, about 68% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation from the mean 50. The z-scores are −1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie within two standard deviations of the mean. Therefore, about 95% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within two standard deviations from the mean 50. The z-scores are −2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie within three standard deviations of the mean. Therefore, about 95% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ from the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The z-scores are −3 and +3 for 32 and 68, respectively.

Your Turn!

From 1984 to 1985, the mean height of 15- to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y represent the height of 15- to 18-year-old males in 1984 to 1985. Then Y ~ N(172.36, 6.34).

a. About 68% of the y values lie between what two values? These values are _______ and _______. The z-scores are _______and _______, respectively.
b. About 95% of the y values lie between what two values? These values are _______ and _______. The z-scores are _______ and _______, respectively.
c. About 99.7% of the y values lie between what two values? These values are _______ and _______. The z-scores are _______ and _______, respectively.

Solution
a. About 68% of the values lie between **166.02 cm** and **178.7 cm**. The z-scores are **–1** and **1**.
b. About 95% of the values lie between **159.68 cm** and **185.04 cm**. The z-scores are **–2** and **2**.
c. About 99.7% of the values lie between **153.34 cm** and **191.38 cm**. The z-scores are **–3** and **3**.

Finding Normal Probabilities

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$.

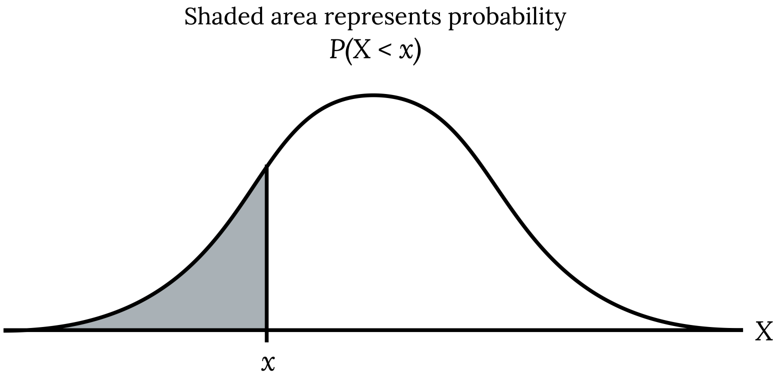


Figure 4.21: P(X < x). *Figure description available at the end of the section*.

If we know the area to the left, we can then use the complement rule to find the area to the right:

$$P(X > x) = 1 - P(X < x)$$

To find the area between two numbers, we can write the equation in terms of a CDF:

$$P(a < X < b) = P(X < b) - P(X < a)$$

Also recall that for continuous distributions:

$$P(X < x) \cong (X \leq x) \ \& \ P(X > x) \cong P(X \geq x)$$

There are three main ways we could find probabilities associated with the normal distribution:

- Complicated math
- The standardizing process
- Technology

If you recall the formula previously presented for the PDF of the normal distribution, you could imagine why it's preferable to avoid involving complicated math if possible.

In order to work around that, there is a process called standardizing that involves z-scores, the standard normal distribution, and tables. Although this tried and true process is now somewhat antiquated, it is a great place to start.

There are many technologies (e.g., calculators and various pieces of statistical software) that let us skip the entire standardizing process and instantaneously provide us with a probability. Although we typically have these at our disposal to use in practice, it is good to understand the process going on behind the scenes to make sure we apply our technology correctly.

The Standard Normal Distribution

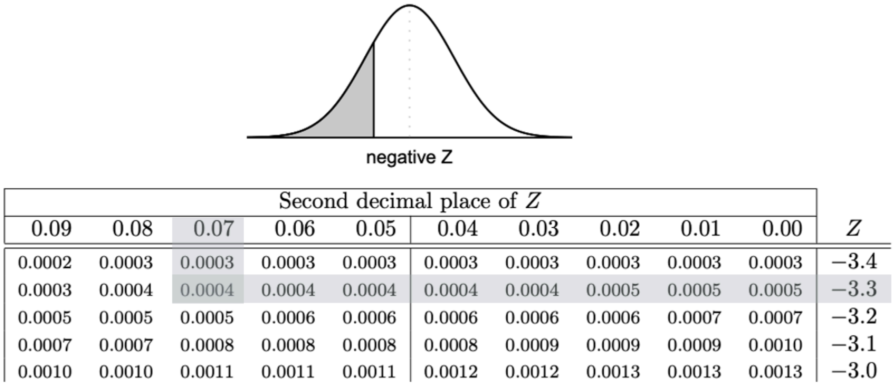
The **standard normal distribution (SND)** is the simplest form of the normal distribution. The mean for the standard normal distribution is zero, and the standard deviation is one. The transformation $z = \frac{x-\mu}{\sigma}$ produces the distribution Z ~ N(0, 1). The value x in the given equation comes from a normal distribution with mean μ and standard deviation σ.

Recall our previous discussion of **z-scores**, which are converted to units of the standard deviation. If X is a normally distributed random variable and X ~ N(μ, σ), then the z-score is:

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

Recall a z-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero.

We have the z-table at our disposal with probabilities already calculated and organized. Note that most z-tables give us the left-tailed, CDF, or "less than" probability. For example, the area to the left of a z-score of -3.37 is P(Z ≤ -3.37) = 0.0004.



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0

Figure 4.22: Z-table. Figure description available at the end of the section.

The SND CDF value, P(Z ≤ z), is also denoted as Φ(z). We can then use these CDF values, P(Z ≤ z), and some probability rules to find greater than [P(Z ≥ z) = 1-P(Z ≤ z)] or in-between [P(a ≤ Z ≤ b) = P(Z ≤ b) − P(Z ≤ a)] probabilities.

Example

Use the z-table to find the following probabilities:

P(Z ≤ 1)

Solution
0.8413

P(Z ≥ 1)

Solution
0.1587

P(-1 ≤ Z ≤ 1)

Solution
0.6826

Your Turn!

Use the z-table to find the following probabilities:

P(Z ≤ -0.54)

P(Z ≥ 1.2)

P(-1.5 ≤ Z ≤ 0.84)

The Standardizing Process

So far, we have discussed converting any normal distribution with any mean and standard deviation to the standard normal distribution in units of *z*-scores. We also have the associated probabilities in our *z*-table. Essentially, the work has been done for us if we know how to standardize and look up the associated probability in the table. The general process is:

X ~ N(μ, σ) -> Z ~ N(0, 1) -> probability from *z*-table

While maybe outdated in our technology age, this process is good for beginners to understand and useful when we do not have access to technology.

Example

Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed.

$$X \sim N(10.2, 0.8)$$

Calculate the z-scores that correspond to the following weights, then find the associated probabilities.

The probability that a child weighs less than 11 kg

Solution
$(11 - 10.2)/0.8 = 1$

A child who weighs 11 kg is one standard deviation above the mean of 10.2 kg.

$P(Z \leq 1) = 0.8413$

The probability that a child weighs more than 7.9 kg

Solution
$(7.9 - 10.2)/0.8 = -2.875$

A child who weighs 7.9 kg is 2.875 standard deviations below the mean of 10.2 kg.

$P(Z \geq -2.88) = 1 - P(Z \leq -2.88) = 1 - 0.002 = 0.998$

The probability that a child weighs between 11.2 and 12.2 kg

Solution
$z1 = (11.2 - 10.2)/0.8 = 1.25$ and $z2 = (12.2 - 10.2)/0.8 = 2.5$

A child who weighs 12.2 kg is 2.5 standard deviation above the mean of 10.2 kg.

$P(1.25 \leq Z \leq 2.5) = P(Z \leq 2.5) - P(Z \leq 1.25) = 0.9938 - 0.8944 = 0.0994$

Your Turn!

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

Find the probability that a randomly selected golfer scored less than 65.

Find the probability that a golfer scored between 66 and 70.

Working Backwards

Sometimes, we may be given a percentile or z-score and want to work backward through the standardizing process to find a value on the original distribution. This “un-standardizing” process of finding a normal **quantile** or percentile associated with the normal distribution looks like this:

Probability in z-table -> Z ~ N(0, 1) -> X ~ N(μ, σ)

For example, if the mean of a normal distribution is five and the standard deviation is two, what value is three standard deviations above (to the right of) the mean (z-score = 3). Rearranging the z-score formula, the calculation is as follows:

$$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$$

Often, we are given a percentile to find on the original distribution. For example, what if we want to know a value on the previous distribution that corresponds to the 90th percentile? We can look up a probability of 0.9 in the z-table and find a corresponding z-score of approximately 1.28.

$$x = \mu + (z)(\sigma) = 5 + (1.28)(2) = 7.56$$

Example

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

Solution
6.16

The middle 20% of mandarin oranges from this farm have diameters between ______ and ______.

Solution
Between 5.79 and 5.91

Your Turn!

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean μ = 81 points and standard deviation σ = 15 points.

a. Calculate the first and third quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 4.19: Kindred Grey (2020). *Normal distribution*. CC BY-SA 4.0.

Figure 4.20: Kindred Grey (2020). *Empirical rule*. CC BY-SA 4.0.

Figure 4.21: Kindred Grey (2020). *P*(X < *x*). CC BY-SA 4.0.

Figure 4.22: Kindred Grey (2020). Z-*table*. CC BY-SA 4.0.

Figure Descriptions

Figure 4.19: Bell-shaped curve diagram with the lower case Greek letter mu at the center of the x axis. It has the label Normal: uppercase X is similar to N (μ, σ)

Figure 4.20: Frequency curve that illustrates the empirical rule. The normal curve is shown over a horizontal axis. The axis is labeled with points -3s, -2s, -1s, m, 1s, 2s, 3s. Vertical lines connect the axis to the curve at each labeled point. The peak of the curve aligns with the point m.

Figure 4.21: Diagram showing a bell-shaped curve with uppercase X at the extreme right end of the X axis. The X axis also contains a lowercase x about one-quarter of the way across the X axis from the right. The area under the bell curve to the right of the lowercase x is shaded. The label states: shaded area represents probability P(X less than x).

Figure 4.22: Z score table that highlights the associated value of 0.0004 with Z value of -3.37.

4.6 The Normal Approximation to the Binomial

The **binomial formula** is cumbersome when the sample size (n) is large, particularly when we consider a range of observations, as shown in the following example.

Example

Approximately 15% of the US population smokes cigarettes. A local government believed their community had a lower smoking rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 42 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?

How to solve this
We first need to verify the four conditions for the binomial model are met. The question posed is equivalent to asking, what is the probability of observing $k = 0, 1, 2, \ldots,$ or 42 smokers in a sample of $n = 400$ when $p = 0.15$? We can compute these 43 different probabilities and add them together.

Solution
$\mathrm{P}(k = 0 \text{ or } k = 1 \text{ or } \cdots \text{ or } k = 42)$
$= \mathrm{P}(k = 0) + \mathrm{P}(k = 1) + \cdots + \mathrm{P}(k = 42)$
$= 0.0054$

The computations in the previous example are tedious, long, and nearly impossible if you do not have access to technology. In some cases, we may use the normal distribution as an easier and faster way to estimate binomial probabilities. In general, we should avoid long, tedious work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. We might wonder if it is reasonable to use the normal model in place of the binomial distribution. Surprisingly, yes, if certain conditions are met.

Binomial Approximation Conditions

Consider the binomial model when the probability of a success is $p = 0.10$. The following figure shows four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: n

= 10, 30, 100, 300. What happens to the shape of the distributions as the sample size increases? What distribution does the last histogram resemble?

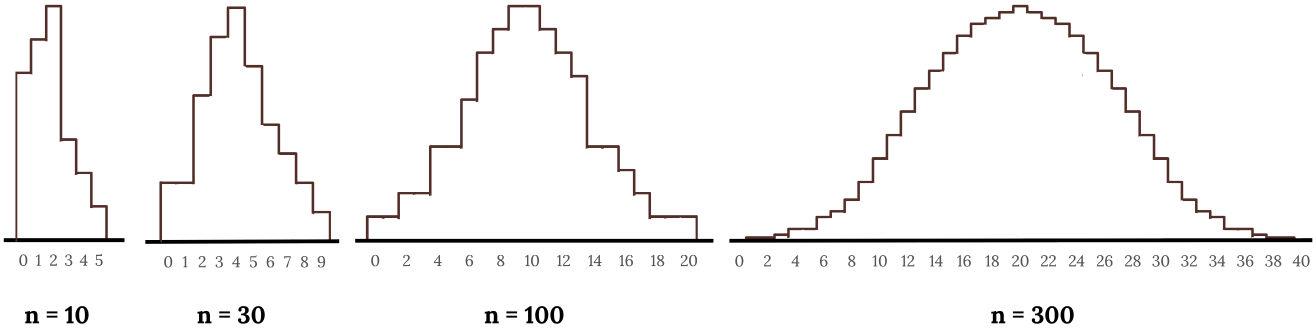


Figure 4.23: Hollow histograms for different sample sizes. *Figure description available at the end of the section.*

By the last histogram, it appears the distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution.

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1 - p)$ are both *at least* 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \text{ and } \sigma = np(1 - p)$$

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting of the previous example.

Example (Continued)

Use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$.

Already aware of the binomial model, we then verify that both np and $n(1 - p)$ are at least 10:

- $np = 400 \times 0.15 = 60$ $n(1 - p) = 400 \times 0.85 = 340$

With these conditions met, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

- $\mu = np = 60$ and $\sigma = np(1 - p) = 7.14$

We want to find the probability of observing 42 or fewer smokers using this model. Use the normal model N($\mu = 60$, $\sigma = 7.14$) and standardize to estimate the probability of observing 42 or fewer smokers. Your answer should be approximately equal to the solution we found in the previous of example, 0.0054.

Compute the z-score first.

Solution
Z = (42−60)/7.14 = −2.52.

The corresponding left tail area from the table or technology is 0.0059.

The Continuity Correction

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when p = 0.15. With such a large sample, we might be tempted to apply the normal approximation and use the range 49–51. However, we would find that the binomial solution and the normal approximation notably differ:

- Binomial: 0.0649
- Normal: 0.0421

We can identify the cause of this discrepancy in the next figure, which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval.

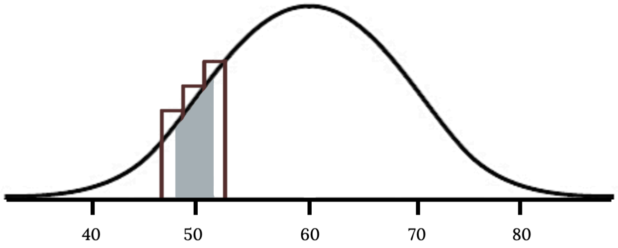


Figure 4.24: Continuity correction. Figure description available at the end of the section.

The normal approximation to the binomial distribution for intervals of values can usually be improved if cutoff values are modified slightly. The cutoff values for the lower end of a shaded region should be reduced by 0.5, and the cutoff value for the upper end should be increased by 0.5. This is called the **continuity correction.**

The tip to add extra area when applying the normal approximation is most often useful when examining a range of observations. In the example above, the revised normal distribution estimate is 0.0633, much closer to the exact value of 0.0649. While it is possible to also apply this correction when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 4.23: Kindred Grey (2020). *Hollow histograms for different sample sizes.* CC BY-SA 4.0.

Figure 4.24: Kindred Grey (2020). *Continuity correction.* CC BY-SA 4.0.

Figure Descriptions

Figure 4.23: Four hollow histograms side by side. First: represents n = 10 and has higher values towards 0-2 and lower ones to the right. Second: represents n = 30 and has higher values around 4 with lower ones to the left and right of four. Third: represents n = 100 and has x values ranging from zero to 20. Follows a bell shape. Fourth: represents n = 300 and has x values ranging from 10-50. Follows a bell shape.

Figure 4.24: A bell shaped curve with x axis ranges from 40-80 by 10. A section of the graph is highlighted on x = 50.

Chapter 4 Wrap-Up

Concept Check

Take this quiz to check your comprehension of this chapter.

If you are using an offline version of this text, access the quiz for this chapter via the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=165#h5p-123

Section Resources

If you are using an offline version of this text, access these materials by visiting https://doi.org/10.7294/26207456.

4.1 Introduction to Probability and Random Variables

4.2 Discrete Random Variables

4.3 The Binomial Distribution

4.4 Continuous Random Variables

4.5 The Normal Distribution

4.6 The Normal Approximation to the Binomial

Key Terms

Try to define the terms below on your own. Check your response by clicking on the term, or looking at the end-of-book glossary!

4.1 Introduction to Probability and Random Variables

- **Probability**
- **Probability experiment**
- **Outcome**
- **Sample space**
- **Event**
- **Probability model**
- **Law of large numbers**
- **Complement**

4.2 Discrete Random Variables

- **Discrete random variable**
- **Probability mass function (PMF)**
- **Cumulative distribution function (CDF)**
- **Expected value**

4.3 The Binomial Distribution

- **Binomial distribution**
- **Independence**
- **Bernoulli trial**

4.4 Continuous Random Variables

- **Continuous random variable (CRV)**
- **Probability density function (PDF)**
- **Uniform distribution**

4.5 The Normal Distribution

- **Normal (Gaussian) distribution**
- **Empirical rule**
- **Standard normal distribution (SND)**
- **z-score**
- **Quantile**

4.6 The Normal Approximation to the Binomial

- **Continuity correction**

Extra Practice

Extra practice problems are available at the end of the book (Chapter 4 Extra Practice).

CHAPTER 5: FOUNDATIONS OF INFERENCE

5.1 Point Estimation and Sampling Distributions

Learning Objectives

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

- Understand point estimation
- Apply and interpret the central limit theorem
- Construct and interpret confidence intervals for means when the population standard deviation is known
- Understand the behavior of confidence intervals
- Carry out hypothesis tests for means when the population standard deviation is known
- Understand the probabilities of error in hypothesis tests

Statistical Inference

It is often necessary to guess, infer, or generalize about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people choose games based on their perceived likelihood of winning. You may have chosen your course of study based on the probable availability of jobs.

Figure 5.1: Loose change. If you want to figure out the distribution of the change people carry in their pockets, and your sample is large enough, you will find that the distribution follows certain patterns. Figure description available at the end of the section.

Statistical inference uses what we know about probability to make our best guesses, or estimates, from samples about the population from which they came. The main forms of inference are:

1. Point estimation
2. Confidence interval
3. Hypothesis testing

Point Estimation

Suppose you are trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a **point estimate** of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempt. In this case, you would have obtained a point estimate for the true proportion.

The most natural way to estimate features of the population (**parameters**) is to use the corresponding summary **statistic** calculated from the sample. Some common point estimates and their corresponding parameters are found in the following table:

Parameter	Measure	Statistic
μ	Mean of a single population	$\bar{x}$
p	Proportion of a single population	$\hat{p}$
μ_D	Mean difference of two dependent populations (matched pairs)	$\bar{x}_D$
$\mu_1 - \mu_2$	Difference in means of two independent populations	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	Difference in proportions of two populations	$\hat{p}_1 - \hat{p}_2$
σ^2	Variance of a single population	S^2
σ	Standard deviation of a single population	S

Figure 5.2: Parameters and point estimates

Suppose the mean weight of a sample of 60 adults is 173.3 lbs; this sample mean is a point estimate of the population mean weight, μ. Remember this is one of many samples that we could have taken from the population. If a different random sample of 60 individuals was taken from the same population, the new sample mean would likely be different as a result of **sampling variability**. While estimates generally vary from one sample to another, the population mean is a fixed value.

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a point estimate of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the parameter of interest. When the parameter is a proportion, it is often denoted by *p*, and we often refer to the sample proportion as $< em > \hat{p} < /em >$ (pronounced "*p*-hat"). Unless we collect responses from every individual in the population, *p* remains unknown, and we use $< em > \hat{p} < /em >$ as our estimate of *p*.

How would one estimate the difference in average weight between men and women? Suppose a sample of men yields a mean of 185.1 lbs, and a sample of women men yields a mean of 162.3 lbs. What is a good point estimate for the difference in these two population means? We will expand on this in following chapters.

Sampling Distributions

We have established that different samples yield different statistics due to sampling variability. These statistics have their own distributions, called sampling distributions, that reflect this as a random variable. The **sampling distribution** of a sample statistic is the distribution of the point estimates based on samples of a fixed size, n, from a certain population. It is useful to think of a particular point estimate as being drawn from a sampling distribution.

Recall the sample mean weight calculated from a previous sample of 173.3 lbs. Suppose another random sample of 60 participants might produce a different value of x, such as 169.5 lbs. Repeated random sampling could result in additional different values, perhaps 172.1 lbs, 168.5 lbs, and so on. Each sample mean can be thought of as a single observation from a random variable X. The distribution of X is called the sampling distribution of the sample mean, and it has its own mean and standard deviation like the random variables discussed previously. We will simulate the concept of a sampling distribution using technology to repeatedly sample, calculate statistics, and graph them. However, the actual sampling distribution would only be attainable if we could theoretically take an infinite amount of samples.

Each of the point estimates in the table above have their own unique sampling distributions that we will explore in the future.

Unbiased Estimation

Although variability in samples is present, there remains a fixed value for any population parameter. What makes a statistical estimate of this parameter of interest “good”? It must be both accurate and precise.

The accuracy of an estimate refers to how well it estimates the actual value of that parameter. Mathematically, this is true when the expected value of your statistic is equal to the value of that parameter. Visually, this looks like the center of the sampling distribution being situated at the value of that parameter.

According to the **law of large numbers**, probabilities converge to what we expect over time. Point estimates follow this rule, becoming more accurate with increasing sample size. The figure below shows the sample mean weight calculated for random samples drawn, where sample size increases by one for each draw until sample size equals 500. The maroon dashed horizontal line is drawn at the average weight of all adults (169.7 lbs), which represents the population mean weight according to the CDC.

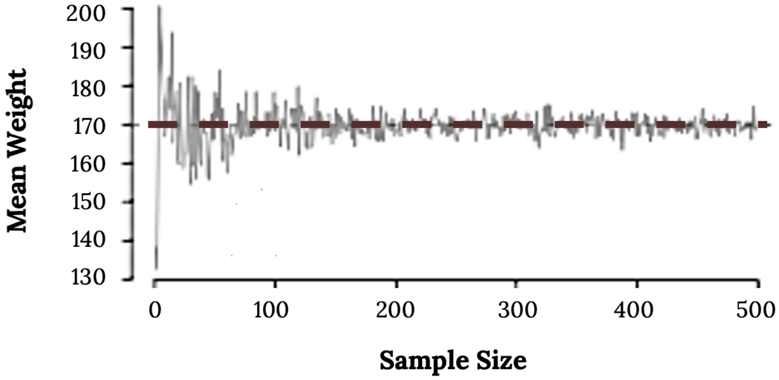


Figure 5.3: Law of large numbers. *Figure description available at the end of the section.*

Note how a sample size around 50 may produce a sample mean that is as much as 10 lbs higher or lower than the population mean. As sample size increases, the fluctuations around the population mean decrease; in other words, as sample size increases, the sample mean becomes less variable and provides a more reliable estimate of the population mean.

In addition to accuracy, a precise estimate is also more useful. This means that the values of the statistics seem pretty close together over repeated sampling. The precision of an estimate can be visualized as the spread of the sampling distribution, usually quantified by the standard deviation. The standard deviation of a sampling distribution is often referred to as the **standard error.** Smaller standard errors are affected by sample size and lead to more precise estimates.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 5.1: Matthew Lancaster (2020). *Silver round coins on clear glass jar.* Unsplash license. https://unsplash.com/photos/silver-round-coins-on-clear-glass-jar-ip1eu-cw49A

Figure 5.3: Kindred Grey (2020). *Law of large numbers*. CC BY-SA 4.0.

Figure Descriptions

Figure 5.1: A glass jar of coins sits on a dark wooden table.

Figure 5.3: Line graph with sample size on the x axis and mean weight on the y axis. When sample size is small (x = 10) the mean weight varies. As sample size increases (x = 500) the mean weight does not vary as much. Overall graph mimics a cornucopia shape with the wide side on the left and narrow side on the right.

5.2 The Sampling Distribution of the Sample Mean (Central Limit Theorem)

Let's start our foray into inference by focusing on the sample mean. Why are we so concerned with means? Two reasons: they give us a middle ground for comparison, and they are easy to calculate. In this section, we will see what we can deduce about the sampling distribution of the sample mean.

The Central Limit Theorem for a Sample Mean

The **central limit theorem (CLT)** is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both forms are concerned with drawing finite samples sizes, n, from a population with a known mean, μ, and a known standard deviation, σ. One of the forms says that, if we collect samples of size n with a "large enough" n, then the resulting distribution can be approximated by the normal distribution.

Applying the law of large numbers here, we could say that taking larger and larger samples from a population brings the mean, $\overline{x}$, of the sample closer and closer to μ. From the central limit theorem, we know that the sample means increasingly follow a normal distribution as n gets larger and larger. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\overline{x}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean, $\overline{x}$, must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

The size of the sample, n, that is considered "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30, or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. Sampling is done with replacement.

The following images look at sampling distributions of the sample mean built from taking 1,000 samples of different sample sizes from a normal population. What pattern do you notice?

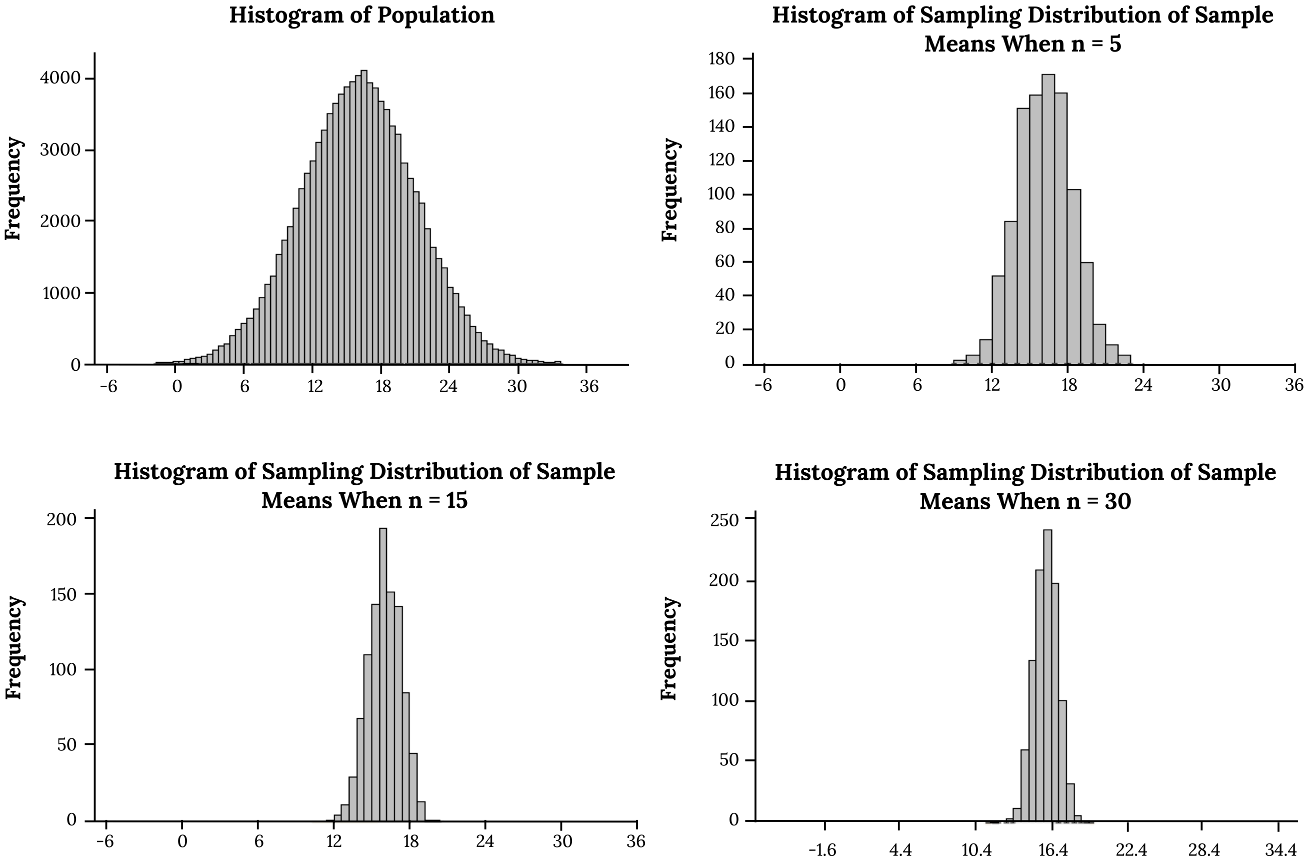


Figure 5.4: Sampling distributions of the sample mean from a normal population. Figure description available at the end of the section.

The following images look at sampling distributions of the sample mean built from taking 1,000 samples of different sample sizes from a non-normal population (in this case, it happens to be exponential). What pattern do you notice?

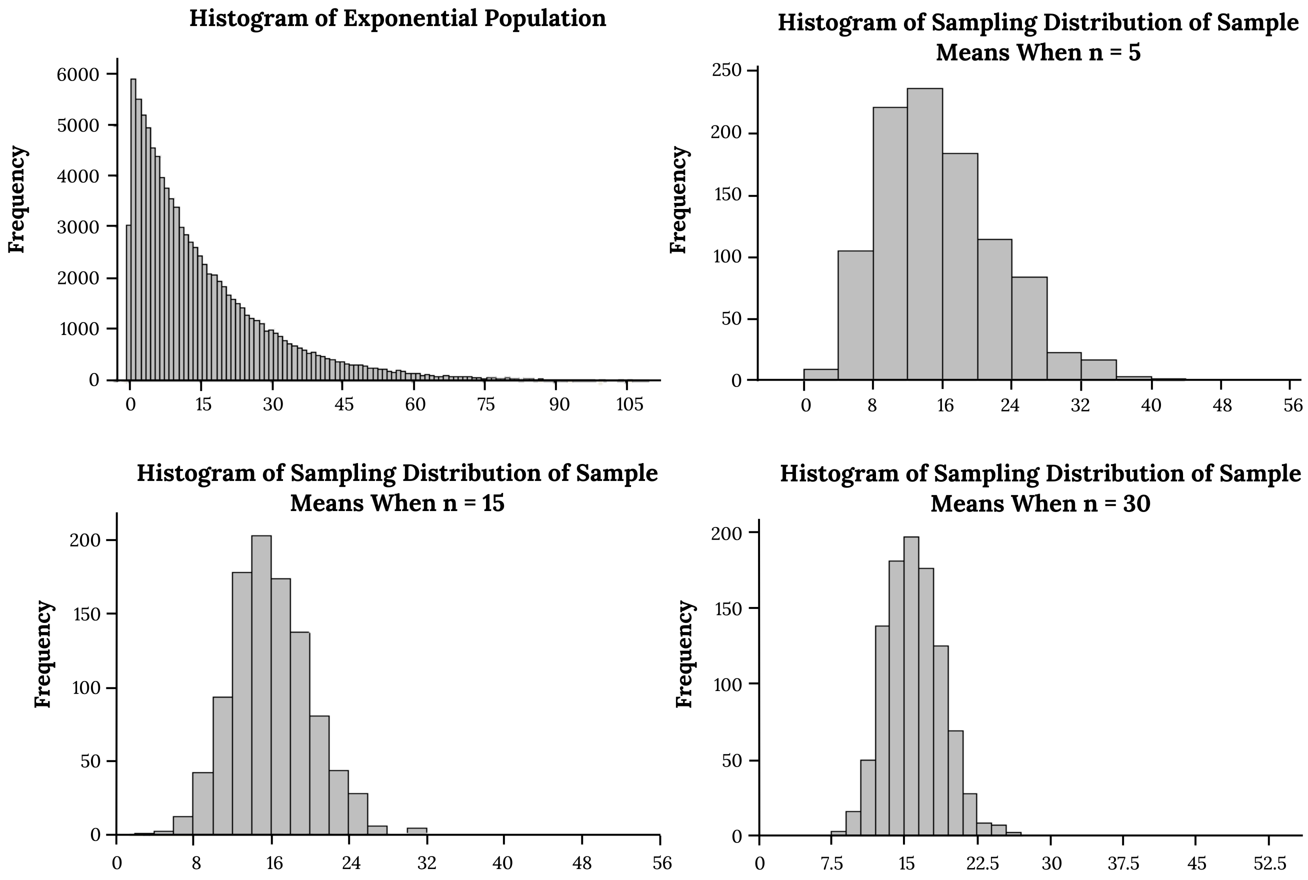


Figure 5.5: Sampling distributions of the sample mean from a non-normal population. *Figure description available at the end of the section*.

What differences do you notice when sampling from normal and non-normal populations?

Example

Suppose:

- eight students roll one fair die ten times
- seven roll two fair dice ten times
- nine roll five fair dice ten times
- 11 roll ten fair dice ten times

Each time a person rolls more than one die, he or she calculates the sample mean of the faces showing. For example, one person might roll five fair dice once and get 2, 2, 3, 4, 6.

The mean is $\frac{2+2+3+4+6}{5}$ = 3.4. The 3.4 is one mean when five fair dice are rolled. Suppose this person then rolls the five dice nine more times and calculates nine more means (for a total of ten means).

As the number of dice rolled increases from one to two to five to ten, the following would happen:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (i.e., the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

We have just demonstrated the idea of central limit theorem (CLT) for means—as you increase the sample size, the sampling distribution of the sample mean tends toward a normal distribution.

To summarize, the central limit theorem for sample means says that, if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and calculating their means, the sample means form their own normal distribution (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by the sample size. Standard deviation is the square root of variance, so the standard deviation of the sampling distribution (a.k.a. standard error) is the standard deviation of the original distribution divided by the square root of n. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

It would be difficult to overstate the importance of the central limit theorem in statistical theory. Knowing that data behaves in a predictable way—even if its distribution is not normal—is a powerful tool. We can simulate this idea using technology.

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, let:

- μ_X = the mean of X
- σ_X = the standard error of X

The standard deviation of $\overline{x}$ is called the standard error of the mean and is written as:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Note here we are assuming we know the population standard deviation.

If you draw random samples of size n, then as n increases, the random variable $\overline{X}$ which consists of sample means, tends to be normally distributed and the following is true:

$$\overline{x} \sim \mathrm{N}\left(\mu_x,\ \frac{\sigma}{\sqrt{n}}\right).$$

To put it more formally, if you draw random samples of size n, the distribution of the random variable $\overline{X}$, which consists of sample means, is called the sampling distribution of the sample mean. The sampling distribution of the mean approaches a normal distribution as the sample size (n) increases.

Using the CLT

It is important to understand when to use the central limit theorem. If you are being asked to find the probability of an individual value, do not use the CLT. Use the distribution of its random variable. However, if you are being asked to find the probability of the mean of a sample, then use the CLT for the mean.

The z-score associated with random variable $\overline{X}$ differs from the score of a single observation. Remember, the mean $\overline{x}$ is the mean of one sample and μ_X is the average, or center, of both X (the original distribution) and $\overline{X}$.

$$z = \frac{\overline{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

We can take a familiar approach, using a z-table and standardizing, or we can use the technology of our choice.

Example

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size n = 25 are drawn randomly from the population.

Find the probability that the sample mean is between 85 and 92. Let X represent one value from the original unknown population.

Solution

The standard error of the mean is $\frac{\sigma_x}{\sqrt{n}}$ = $\frac{15}{\sqrt{25}}$ = 3. Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size n.

Let $\overline{x}$ = the mean of a sample of size 25. Since μ_X = 90, σ_X = 15, and n = 25, $\overline{x} \sim$ N(90, $\frac{15}{\sqrt{25}}$).

Find P(85 < $\overline{x}$ < 92). Draw a graph.

Solution
This is a "between" problem. You will need to find two z scores, their corresponding probabilities, and then subtract.

$Z_1 = \frac{85-90}{\frac{15}{\sqrt{25}}} = -1.67$

$Z_2 = \frac{92-90}{\frac{15}{\sqrt{25}}} = 0.67$

The probability that the sample mean is between 85 and 92 is 0.7475 – 0.0478 = 0.6997. Check this using technology.

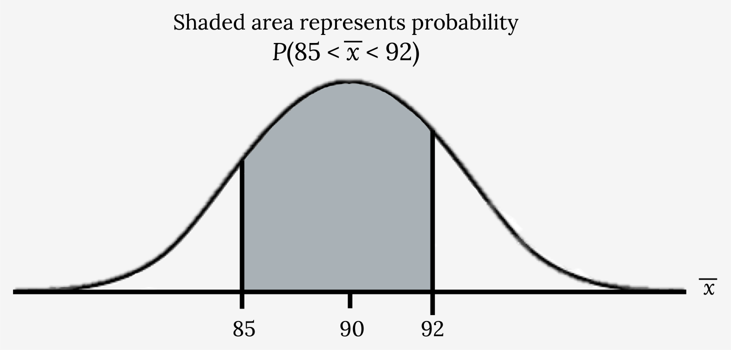


Figure 5.6: Area under the curve. Figure description available at the end of the section.

Find the value that is two standard deviations above 90, the expected value of the sample mean.

Solution
To find the value that is two standard deviations above the expected value 90, use the formula value = μ_X + (#ofTSDEVs)($\frac{\sigma_x}{\sqrt{n}}$)

Value = 90 + 2($\frac{15}{\sqrt{25}}$) = 96

The value that is two standard deviations above the expected value is 96.

Your Turn!

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size n = 30 are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 5.4: Kindred Grey (2021). *Sampling distributions of the sample mean from a normal population.* CC BY-SA 4.0.

Figure 5.5: Kindred Grey (2021). *Sampling distributions of the sample mean from a non-normal population.* CC BY-SA 4.0.

Figure 5.6: Kindred Grey (2020). *Area under the curve.* CC BY-SA 4.0. Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/5-practice

Figure Descriptions

Figure 5.4: Four histograms. Top left—Histogram of Population; Top right—Histogram of Sampling Distribution of Sample Means when n = five; Bottom left—Histogram of Sampling Distribution of Sample Means when n = 15; Bottom right—Histogram of Sampling Distribution of Sample Means when n = 30. All histograms follow typical bell-curve shape and as n increases, the shape gets more narrow around the mean.

Figure 5.5: Four histograms. Top left—Histogram of exponential population; Top right—Histogram of Sampling Distribution of Sample Means when n = five; Bottom left—Histogram of Sampling Distribution of Sample Means when n = 15; Bottom right—Histogram of Sampling Distribution of Sample Means when n = 30. All histograms are skewed right and as n increases, the plot gets more narrow around the mean.

Figure 5.6: Normal distribution curve where the peak of the curve coincides with the point 90 on the horizontal axis. The points 85 and 92 are labeled on the axis. Vertical lines are drawn from these points to the

curve and the area between the lines is shaded. The shaded region represents the probability that 85 < x < 92.

5.3 Introduction to Confidence Intervals

We use **inferential statistics** to make generalizations about an unknown population. The simplest way of doing this is to use the sample data in making a point estimate of a population parameter. We realize that due to **sampling variability**, the point estimate is most likely not the exact value of the population parameter, though it should be close. After calculating point estimates, we can build off of them to construct interval estimates called confidence intervals.

Figure 5.7: M&Ms. Have you ever wondered about the average number of M&Ms in a bag at the grocery store? You can use confidence intervals to answer this question. Figure description available at the end of the section.

Confidence Intervals

A **confidence interval** is another type of estimate, but, instead of being just one number, it is a range of reasonable values in which we expect the population parameter to fall. Since a point estimate may not be perfect due to variability, the idea is to build an interval based on a point estimate to hopefully capture the parameter of interest in the interval. There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success. It is important to keep in mind that the confidence interval itself is a random variable, while the population parameter is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\overline{x}$. You would use $\overline{x}$ to estimate the population mean. The sample mean, $\overline{x}$, is the point estimate for the population mean, μ.

Continuing the iTunes example, suppose we do not know the population mean, μ, but we do know that the population standard deviation is $\sigma = 1$ and that our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is:

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **empirical rule**, which applies to bell-shaped distributions, says that the sample mean, $\overline{x}$, will be within two standard deviations of the population mean, μ, in approximately 95% of the samples. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean is likely to be within 0.2 units of μ.

Because $\overline{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\overline{x}$ in 95% of the samples. The population mean, μ, is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\overline{x}$ − 0.2 and $\overline{x}$ + 0.2 in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\overline{x}$ = 2. Therefore, the unknown population mean μ is between $\overline{x}$ − 0.2 = 2 − 0.2 = 1.8 and $\overline{x}$ + 0.2 = 2 + 0.2 = 2.2.

We can say that we are about 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The approximate 95% confidence interval is (1.8, 2.2).

This approximate 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ, or our sample produced an $\overline{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that confidence intervals are created for an unknown population parameter. Confidence intervals for most parameters have the form:

(Point Estimate ± Margin of Error) = (Point Estimate − Margin of Error, Point Estimate + Margin of Error)

The **margin of error (MoE)** depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. If our sample has a mean of $\overline{x}$ = 10, we can construct the 90% confidence interval (5, 15) where MoE = 5.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean, μ, where the population standard deviation is known, we need $\overline{x}$ as an estimate for μ and we need the margin of error. The sample mean, $\overline{x}$, is the point estimate of the unknown population mean, μ.

A confidence interval estimate will have the following form:

PE−MoE, PE+MoE

As a result, a confidence interval for the unknown population mean μ in symbols would look like:

$\overline{x}$ − MoE, $\overline{x}$ + MoE

Remember, the margin of error depends mainly on the confidence level (CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is up to the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of their conclusions.

There is another probability called alpha (α), which is related to the confidence level and represents the chance that the interval does not contain the unknown population parameter. Mathematically, this looks like:

$$\alpha + \text{CL} = 1$$

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

1. Calculate the sample mean, $\overline{x}$, from the sample data. Remember, in this section, we already know the population standard deviation, σ.
2. Find the z-score (critical value) that corresponds to the confidence level.
3. Calculate the margin of error.
4. Construct the confidence interval.
5. Write a sentence that interprets the estimate in the context of the situation in the problem. (Use the words of the problem to explain what the confidence interval means.)

We will first examine each step in more detail and then illustrate the process with some examples.

Example

Suppose we have collected data from a sample. We know the sample mean, but we do not know the mean for the entire population. The sample mean is 7, and the margin of error for the mean is 2.5. Find the confidence interval and interpret.

Solution
$\overline{x} = 7$

Margin of error = 2.5

The confidence interval is (7 – 2.5, 7 + 2.5), and calculating the values gives (4.5, 9.5).

If the confidence level is 95%, then we say that, “We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5.”

Your Turn!

Suppose we have data from a sample. The sample mean is 15, and the margin of error for the mean is 3.2. What is the confidence interval estimate for the population mean?

Changing the Confidence Level

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\overline{x}$ = 10, and we have constructed the 90% confidence interval (5, 15) where the MoE = 5.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the normal distribution.

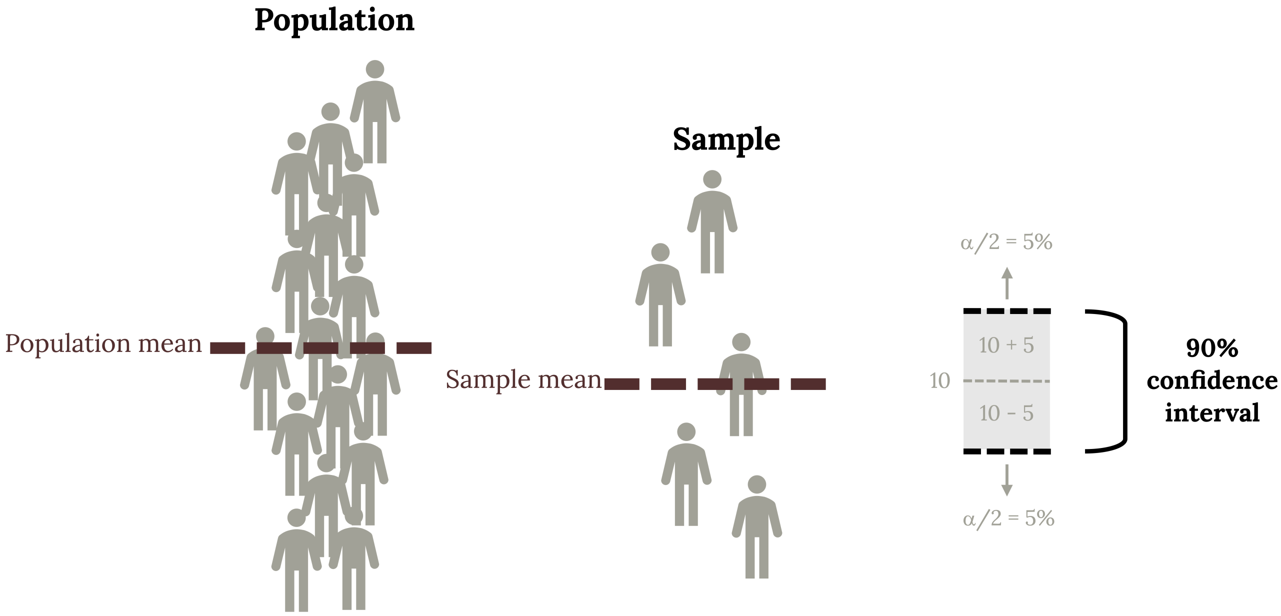


Figure 5.8: 90% confidence level. Figure description available at the end of the section.

To capture the central 90%, we must go out 1.645 standard deviations on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that results in an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the standard deviation used must be appropriate for the parameter we are estimating, so in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$, is commonly called the "standard error of the mean" in order to distinguish clearly the standard deviation for a mean from the population standard deviation, σ.

In summary, as a result of the central limit theorem:

- $\overline{X}$ is normally distributed; that is, $\overline{X}$ ~ N(μ_X, $\frac{\sigma}{\sqrt{n}}$).
- When the population standard deviation, σ, is known, we use a normal distribution to calculate the margin of error.

Finding the Critical Value

When we know the population standard deviation, we use a standard normal distribution to calculate the margin of error and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution Z ~ N(0, 1). This z-score is also called a **critical value**.

The confidence level is the area in the middle of the standard normal distribution. Since CL = 1 − α, α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

NOTE: Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$.

Example

Find the critical value for a 95% confidence interval.

Solution
CL = 0.95, α = 0.05, and α/2 = 0.025; we write $z_{\alpha/2} = z_{0.025}$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 1 − 0.025 = 0.975.

$z_{\alpha/2} = z_{0.025}$ = 1.96, using technology or a standard normal probability table.

Your Turn!

Find the critical value for a 90% confidence interval.

Calculating the Margin of Error

The margin of error formula for an unknown population mean and a known population standard deviation is as follows:

$$\text{MoE} = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

Constructing the Confidence Interval

A confidence interval estimate has the format $\overline{x}$ − MoE, $\overline{x}$ + MoE.

The graph gives a picture of the entire situation.

$\text{CL} + \frac{\alpha}{2} + \frac{\alpha}{2} = \text{CL} + \alpha = 1$.

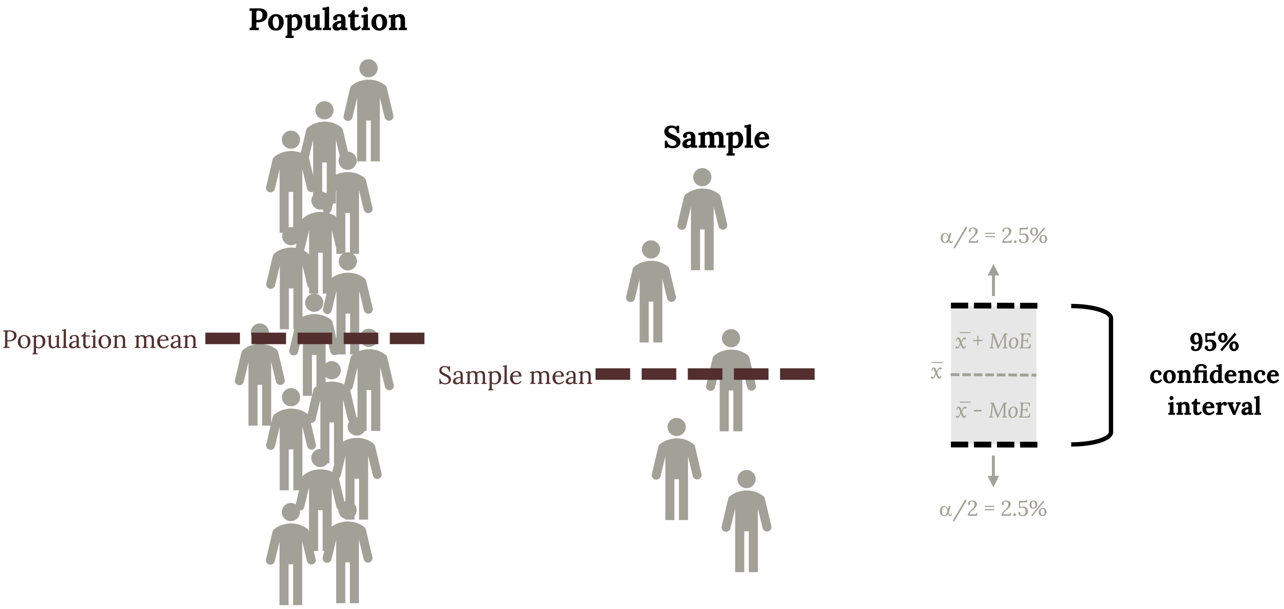


Figure 5.9: Constructing the confidence interval. Figure description available at the end of the section.

Writing the Interpretation

The interpretation should clearly state the confidence level, explain what population parameter is being estimated (here, the population mean), and state the confidence interval (both endpoints). “We can be ____% confident that the interval we created, ______ to _____ , captures the true population mean.” It should include the context of the problem and appropriate units.

Be careful that you do not associate the confidence level with the parameter itself. Your parameter is a fixed value; what is changing is the sample you take and the interval you calculate. We always want to associate the CL% with the sampling process and the interval.

Example

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean score of 68. Find a confidence interval estimate for the population mean exam score (i.e., the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

The step-by-step solution is shown below. If you are comfortable using software, you can use it to calculate the confidence interval directly.

Solution
To find the confidence interval, you need the sample mean, $\overline{x}$, and the margin of error.

$\overline{x}$ = 68

Margin of error = $(z_{\alpha/2})(\frac{\sigma}{\sqrt{n}})$

σ = 3; n = 36; The confidence level is 90% (CL = 0.90)

CL = 0.90 so α = 1 – CL = 1 – 0.90 = 0.10

$\alpha/2$ = 0.05 $z_{\alpha/2}$= $z_{0.05}$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 1 – 0.05 = 0.95.

$z_{\alpha/2}$ = $z_{0.05}$ = 1.645

Margin of error = $(1.645)(\frac{3}{\sqrt{36}})$ = 0.8225

$\overline{x}$ – margin of error = 68 – 0.8225 = 67.1775

$\overline{x}$ + margin of error = 68 + 0.8225 = 68.8225

The 90% confidence interval is **(67.1775, 68.8225)**.

Interpretation: We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Your Turn!

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Find a 90% confidence interval estimate for the population mean delivery time and interpret.

Solution
(34.1347, 37.8653)

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 5.7: Sebastian Gomez (2020). *yellow green and red candies on white ceramic round plate*. Unsplash license. https://unsplash.com/photos/w9pT3v9z1CM

Figure 5.8: Kindred Grey (2024). *90% confidence level.* CC BY-SA 4.0.

Figure 5.9: Kindred Grey (2024). *Constructing the confidence interval.* CC BY-SA 4.0.

Figure Descriptions

Figure 5.7: White bowl with lots of M&Ms sits on a white table.

Figure 5.8: Population is larger than the sample. Population mean is slightly above the sample mean. The sample mean is 10 and the margin of error is 10+5 and 10-5. 90% confidence interval: To 90%, the mean value of the population is in the range 5, 15.

Figure 5.9: Population is larger than the sample. Population mean is slightly above the sample mean. 95% confidence interval: To 95%, the mean value of the population is in this range (x bar + MOE, x bar – MOE).

5.4 The Behavior of Confidence Intervals

Once we know the basics of how to calculate a confidence interval, we also need to know how they behave. In other words, how does tweaking certain parts of the equation effect the interval? Keep in mind that one of the criteria that makes something a "good" statistical estimate is precision. A smaller, or more narrow, interval gives us a more precise and therefore useful estimate.

Changing the Confidence Level or Sample Size

Example

Recall the previous example:

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean score of 68. Find a confidence interval estimate for the population mean exam score (i.e., the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

The 90% confidence interval is (67.1775, 68.8225).

Suppose we change the original problem by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

To find the confidence interval, you need $\overline{x}$, the sample mean, and the MoE.

- $\overline{x}$ = 68
- σ = 3; n = 36; the confidence level is 95% (CL = 0.95)
- MoE = $(z_{\frac{\alpha}{2}})\left(\frac{\sigma}{\sqrt{n}}\right)$

Since CL = 0.95, then α = 1 – CL = 1 – 0.95 = 0.05.

$$\frac{\alpha}{2} = 0.025$$

$$z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025, and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

$$\text{MoE} = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\overline{x} - \text{MoE} = 68 - 0.98 = 67.02$$

$$\overline{x} + \text{MoE} = 68 + 0.98 = 68.98$$

Notice that the MoE is *larger* for a 95% confidence level in the original problem, creating a less precise interval.

Interpretation: We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Alternative Interpretation

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score. Let's compare the results:

The 90% confidence interval is (67.18, 68.82), and the 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the figure below, you'll see that the area 0.99 is larger than the area 0.90, so it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, it is necessary for the confidence interval to be wider.

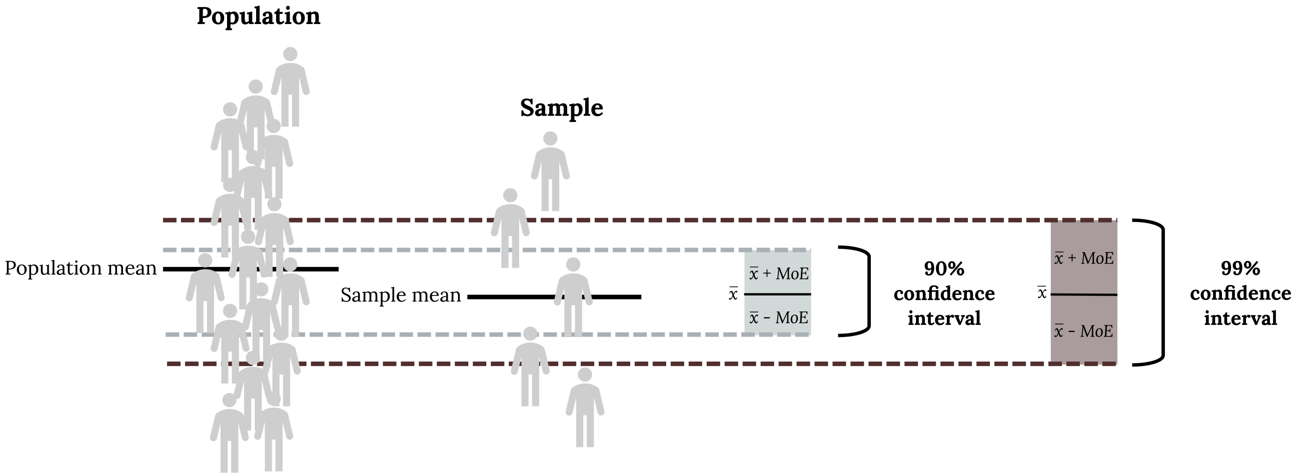


Figure 5.10: Confidence level comparisons. Figure description available at the end of the section.

In conclusion, increasing the confidence level increases the margin of error, making the confidence interval wider.

Working Backwards to Find the Margin of Error or Sample Mean

When we calculate a confidence interval, we must first find the sample mean and calculate the margin of error. However, statistical studies may sometimes state only the confidence interval. If we know the confidence interval, we can work backward to find both the margin of error and the sample mean.

Finding the Margin of Error

- From the upper value for the interval, subtract the sample mean, or
- From the upper value for the interval, subtract the lower value, then divide the difference by two.

Finding the Sample Mean

- Subtract the margin of error from the upper value of the confidence interval, or
- Average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example

Suppose we know that a confidence interval is (67.18, 68.82), and we want to find the margin of error. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the margin of error:

- If we know that the sample mean is 68: MoE = 68.82 − 68 = 0.82.
- If we don't know the sample mean: MoE = $\frac{(68.82-67.18)}{2}$ = 0.82.

Calculate the sample mean:

- If we know the margin of error: $\overline{x}$ = 68.82 − 0.82 = 68
- If we don't know the margin of error: $\overline{x}$ = $\frac{(67.18+68.82)}{2}$ = 68.

Your Turn!

Suppose we know that a confidence interval is (42.12, 47.88). Find the margin of error and the sample mean.

Calculating the Sample Size Needed

If researchers desire a specific margin of error, then they can use the margin of error formula to calculate the required sample size.

Recall the margin of error formula for a population mean when the population standard deviation is known:

$$\text{MoE} = (z_{\frac{\alpha}{2}})\ (\frac{\sigma}{\sqrt{n}})$$

The formula for sample size is $n = \frac{z^2\sigma^2}{MoE^2}$, found by solving the margin of error formula for n.

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and margin of error can use this formula to calculate the size of the sample needed for the study.

Example

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and MoE = 2.
- $z = z_{0.025} = 1.96$ because the confidence level is 95%.
- Use the sample size equation: $n = \frac{z^2\sigma^2}{MoE^2} = \frac{(1.96)^2(15)^2}{2^2} = 216.09$
- Use $n = 217$. (Always round the answer up to the next higher integer to ensure that the sample size is large enough.)

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Example

The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the US falls between $69,720 and $69,922.[1] Find the point estimate for mean US household income and the margin of error for mean US household income.

The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your university to within one inch with 93% confidence. How many male students must you measure?

Use the formula for MoE, solved for *n*:

$$\text{n} = \frac{z^2\sigma^2}{MoE^2}$$

From the statement of the problem, you know that σ = 2.5, and you need MoE = 1.

$$z = z_{0.035} = 1.812$$

(This is the value of *z* for which the area under the density curve to the *right* of *z* is 0.035.)

$$\text{n} = \frac{z^2\sigma^2}{MoE^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$$

You need to measure at least 21 male students to achieve your goal.

Your Turn!

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 5.10: Kindred Grey (2024). *Confidence level comparisons*. CC BY-SA 4.0.

Figure Descriptions

Figure 5.10: Population is larger than the sample. Population mean is slightly above the sample mean. 90% confidence interval: To 90%, the mean value of the population is in this range (x bar + MOE, x bar – MOE). 99% confidence interval has the same formula for finding the range, but is a much wider range. Both intervals include the true population mean, but the 90% interval just barely includes it.

Notes

1. American Fact Finder." U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (accessed July 2, 2013).

5.5 Introduction to Hypothesis Tests

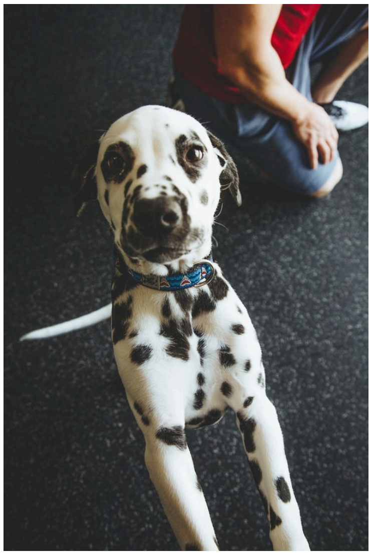


Figure 5.11: Dalmatian spots. You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. [Figure description available at the end of the section](#).

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter.

Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealership advertises that its new small truck gets 35 miles per gallon on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that female managers in their company earn an average of $60,000 per year. A statistician may want to make a decision about or evaluate these claims. A **hypothesis test** can be used to do this.

A hypothesis test involves collecting data from a sample and evaluating the data. Then the statistician makes a decision as to whether or not there is sufficient evidence to reject the null hypothesis based upon analyses of the data.

In this section, you will conduct hypothesis tests on single means when the population standard deviation is known.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will perform some variation of these steps:

1. Define hypotheses.
2. Collect and/or use the sample data to determine the correct distribution to use.
3. Calculate test statistic.
4. Make a decision.
5. Write a conclusion.

Defining your hypotheses

The actual test begins by considering two hypotheses: the null hypothesis and the alternative hypothesis. These hypotheses contain opposing viewpoints.

The **null hypothesis (H_0)** is often a statement of the accepted historical value or norm. This is your starting point that you must assume from the beginning in order to show an effect exists.

The **alternative hypothesis (H_a)** is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a decision**.** There are two options for a decision. They are "reject H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

The following table shows mathematical symbols used in H_0 and H_a:

H_0	H_a
Equal (=)	Not equal (≠) **or** greater than (>) **or** less than (<)
Equal (=)	Less than (<)
Equal (=)	More than (>)

Figure 5.12: Null and alternative hypotheses

NOTE: H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol in the alternative hypothesis depends on the wording of the hypothesis test. Despite this, many researchers may use =, ≤, or ≥ in the null hypothesis. This practice is acceptable because our only decision is to reject or not reject the null hypothesis.

Example

We want to test whether the mean GPA of students in American colleges is 2.0 (out of 4.0). The null hypothesis is: H_0: $\mu = 2.0$. What is the alternative hypothesis?

Solution

H_a: $\mu \neq 2.0$

Your Turn!

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Using the Sample to Test the Null Hypothesis

Once you have defined your hypotheses, the next step in the process is to collect sample data. In a classroom context, the data or summary statistics will usually be given to you.

Then you will have to determine the correct distribution to perform the hypothesis test, given the assumptions you are able to make about the situation. Right now, we are demonstrating these ideas in a test for a mean when the population standard deviation is known using the *z* distribution. We will see other scenarios in the future.

Calculating a Test Statistic

Next you will start evaluating the data. This begins with calculating your **test statistic**, which is a measure of the distance between what you observed and what you are assuming to be true. In this context, your test statistic, z_o, quantifies the number of standard deviations between the sample mean, $\overline{x}$, and the population mean, μ. Calculating the test statistic is analogous to the previously discussed process of standardizing observations with *z*-scores:

$$z = \frac{\overline{x} - \mu_o}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

where μ_o is the value assumed to be true in the null hypothesis.

Making a Decision

Once you have your test statistic, there are two methods to use it to make your decision:

1. Critical value method (discussed further in later chapters)
2. *p*-value method (our current focus)

p-Value Method

To find a ***p*-value**, we use the test statistic to calculate the actual probability of getting the test result. Formally, the *p*-value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.

A large p-value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p-value, the more unlikely the outcome and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example

Suppose a baker claims that his bread height is more than 15 cm on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes ten loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm and the distribution of heights is normal.

The null hypothesis could be H_0: $\mu \leq 15$.

The alternate hypothesis is H_a: $\mu > 15$.

The words "is more than" calls for the use of the > symbol, so "$\mu > 15$″ goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ is known (σ = 0.5 cm), the distribution for the population is known to be normal with mean μ = 15 and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then, is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking how unlikely the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

This means that the p-value is the probability that a sample mean is the same or greater than 17 cm when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

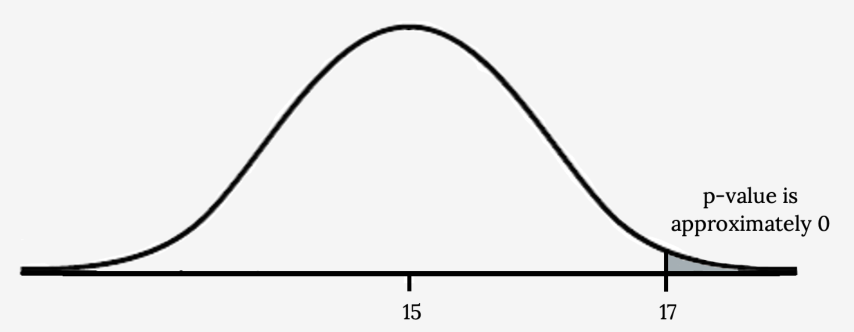


Figure 5.13: Bread height probability. Figure description available at the end of the section.

The p-value is $P(\overline{x} > 17)$, which is approximately zero.

A p-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm purely by CHANCE had the population mean height really been 15 cm. Because the outcome of 17 cm is so unlikely (meaning it is happening NOT by chance alone), we conclude that the evidence is strongly against the null hypothesis that the mean height would be at most 15 cm. There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Your Turn!

A normal distribution has a standard deviation of one. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

Find the p-value.

Solution

H_0: $\mu \leq 12$

H_a: $\mu > 12$

The p-value is 0.0013.

Draw a graph that shows the p-value.

Decision and Conclusion

A systematic way to decide whether to reject or not reject the null hypothesis is to compare the p-value and a preset or preconceived α (also called a **significance level**). A preset α is the probability of a type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem. If there is no given preconceived α, then use $\alpha = 0.05$.

When you make a decision to reject or not reject H_0, do as follows:

- If $\alpha > p$-value, reject H_0. The results of the sample data are **statistically significant**. You can say there is sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis, H_a, may be correct.
- If $\alpha \leq p$-value, fail to reject H_0. The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, H_a, may be correct.

After you make your decision, write a thoughtful conclusion in the context of the scenario incorporating the hypotheses.

NOTE: When you "do not reject H_0," it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_o.

Example

When using the p-value to evaluate a hypothesis test, the following rhymes can come in handy:

If the p-value is low, the null must go.

If the p-value is high, the null must fly.

This memory aid relates a p-value less than the established alpha ("the p-value is low") as rejecting the null hypothesis and, likewise, relates a p-value higher than the established alpha ("the p-value is high") as not rejecting the null hypothesis.

Fill in the blanks:

- Reject the null hypothesis when ______.
- The results of the sample data ______.
- Do not reject the null when hypothesis when ______.
- The results of the sample data ______.

Solution

- Reject the null hypothesis when **the p-value is less than the established alpha value**.
- The results of the sample data **support the alternative hypothesis**.
- Do not reject the null hypothesis when **the p-value is greater than the established alpha value**.
- The results of the sample data **do not support the alternative hypothesis**.

Your Turn!

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

- H_0: p = 0.50, H_a: p > 0.50
- α = 0.01
- p-value = 0.025

Interpret the results and state a conclusion in simple, non-technical terms.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 5.11: Alora Griffiths (2019). *dalmatian puppy near man in blue shorts kneeling*. Unsplash license. https://unsplash.com/photos/7aRQZtLsvqw

Figure 5.13: Kindred Grey (2020). *Bread height probability*. CC BY-SA 4.0.

Figure Descriptions

Figure 5.11: Dalmatian puppy near man sitting on the floor.

Figure 5.13: Normal distribution curve on average bread heights with values 15, as the population mean, and 17, as the point to determine the p-value, on the x-axis.

5.6 Hypothesis Tests in Depth

Establishing the parameter of interest, type of distribution to use, the test statistic, and p-value can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when interpreting the results.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very unlikely to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. Remember that your assumption is just an assumption; it is not a fact, and it may or may not be true. But your sample data are real and are showing you a fact that seems to contradict your assumption.

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside. There are 200 plastic bubbles in the basket, and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200}$ = 0.005. Because this is so unlikely, Ali is hoping they had been misinformed and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Errors in Hypothesis Tests

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

	H_0 is *actually*	
Action	**True**	**False**
Do not reject H_0	Correct outcome	Type II error
Reject H_0	Type I error	Correct outcome

Figure 5.14: Type I and type II errors

The four possible outcomes in the table are:

1. The decision is not to reject H_0 when H_0 is true (correct decision).
2. The decision is to reject H_0 when H_0 is true (incorrect decision known as a **type I error**).
3. The decision is not to reject H_0 when, in fact, H_0 is false (incorrect decision known as a **type II error**).
4. The decision is to reject H_0 when H_0 is false (correct decision whose probability is called the **power** of the test).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a type I error = P(type I error) = probability of rejecting the null hypothesis when the null hypothesis is true. These are also known as false positives. We know that α is often determined in advance, and α = 0.05 is often widely accepted. In that case, you are saying, "We are OK making this type of error in 5% of samples." In fact, the p-value is the exact probability of a type I error based on what you observed.

β = probability of a type II error = P(type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false. These are also known as false negatives.

The **power** of a test is 1 – β.

Ideally, α and β should be as small as possible because they are probabilities of errors but are rarely zero. We want a high power that is as close to one as well. Increasing the sample size can help us achieve these by reducing both α and β and therefore increasing the power of the test.

Example

Suppose the null hypothesis, H_0, is that Frank's rock climbing equipment is safe.

Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. Type II error: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = probability that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β = probability that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the type II error, in which Frank thinks his rock climbing equipment is safe, so he goes ahead and uses it.

Your Turn!

Suppose the null hypothesis, H_0, is that the blood cultures contain no traces of pathogen X. State the type I and type II errors.

Statistical Significance vs. Practical Significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample. Sometimes, researchers will take such large samples that even the slightest difference is detected, even differences where there is no practical value. In such cases, we still say the difference is **statistically significant**, but it is not practically significant.

For example, an online experiment might identify that placing additional ads on a movie review website statistically significantly increases viewership of a TV show by 0.001%, but this increase might not have any practical value.

One role of a data scientist in conducting a study often includes planning the size of the study. The data scientist might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain other information, such as a very rough estimate of the true proportion p, so that she could roughly estimate the standard error. From here, she could suggest a sample size that is sufficiently large enough to detect the real difference if it is meaningful. While larger sample sizes may still be used, these calculations are especially helpful when considering costs or potential risks, such as possible health impacts to volunteers in a medical study.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Chapter 5 Wrap-Up

Concept Check

Take this quiz to check your comprehension of this chapter.

If you are using an offline version of this text, access the quiz for this chapter via the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=286#h5p-154

Section Resources

If you are using an offline version of this text, access these materials by visiting https://doi.org/10.7294/26207456.

5.1 Point Estimation and Sampling Distributions

5.2 The Sampling Distribution of the Sample Mean (CLT)

5.3 Introduction to Confidence Intervals

5.4 The Behavior of Confidence Intervals

5.5 Introduction to Hypothesis Tests

5.6 Hypothesis Tests in Depth

Key Terms

Try to define the terms below on your own. Check your response by clicking on the term, or looking at the end-of-book glossary!

5.1 Point Estimation and Sampling Distributions

- **Statistical inference**
- **Point estimate**
- **Parameter**
- **Statistic**
- **Sampling variability**
- **Sampling distribution**
- **Law of large numbers**
- **Standard error**

5.2 The Sampling Distribution of the Sample Mean (CLT)

- **Central limit theorem (CLT)**

5.3 Introduction to Confidence Intervals

- **Inferential statistics**
- **Confidence interval**
- **Empirical rule**
- **Margin of error (MoE)**
- **Critical value**

5.4 The Behavior of Confidence Intervals

5.5 Introduction to Hypothesis Tests

- **Hypothesis test**
- **Null hypothesis (H_0)**
- **Alternative hypothesis (H_A)**
- **Test statistic**
- ***p*-value**
- **Significance level**
- **Statistically significant**

5.6 Hypothesis Tests in Depth

- **Type I error**

- **Type II error**
- **Power**

Extra Practice

Extra practice problems are available at the end of the book (Chapter 5 Extra Practice).

CHAPTER 6: INFERENCE FOR ONE SAMPLE

6.1 The Sampling Distribution of the Sample Mean (*t*)

Learning Objectives

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

- Construct and interpret confidence intervals for means when the population standard deviation is unknown
- Carry out hypothesis tests for means when the population standard deviation is unknown
- Construct and interpret confidence intervals for a proportion
- Understand the behavior of confidence intervals for a proportion
- Carry out hypothesis tests for a proportion

Figure 6.1: William Gosset (Student). William Sealy Gosset wrote under the pseudonym "Student" so that readers would not know he was a scientist at Guinness Brewery. Figure description available at the end of the section.

We have discussed the **sampling distribution** of the sample mean when the population standard deviation, σ, is known. However, in practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation, s, as an estimate for σ and proceeded as before, calculating a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size can cause inaccuracies in the confidence interval.

Student's *t*-Distribution

William S. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland, ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not always produce accurate results when he tried to use existing inference techniques. He realized that he could not use a normal distribution for the calculation since finding that the actual distribution depends on the sample size. This is because s is a more reliable estimate of σ as samples get bigger. This problem led him to "discover" what is called **Student's *t*-distribution** (after Gosset's pen name, Student).

Until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and used the Student's *t*-distribution only for sample sizes of at most 30. In our current age of technology, the oft-accepted practice now is to simply use the Student's *t*-distribution whenever s is used as an estimate for σ.

In summary, if you draw a simple random sample of size *n* from a population that has an approximately normal distribution with mean μ and unknown population standard deviation σ and calculate the *t*-score, $t = \frac{\overline{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the *t*-scores follow a Student's *t*-distribution with *n* – 1 degrees of freedom. The *t*-score has the same interpretation as the *z*-score. It measures how far $\overline{x}$ is from its mean, μ. For each sample size *n*, there is a different Student's *t*-distribution.

The following images compare the *z* (standard normal) and *t* (Student's *t*). What differences do you notice?

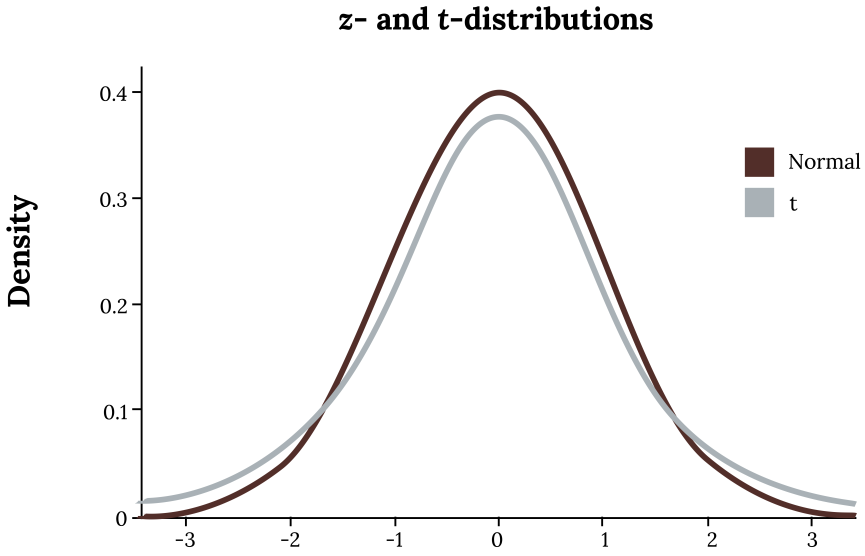


Figure 6.2: Comparing the standard normal distribution and Student's t-distribution. Figure description available at the end of the section.

Degrees of Freedom

The **degrees of freedom (df)**, come from the calculation of the sample standard deviation, s. Remember when we calculated a sample standard deviation, we divided the sum of the squared deviations by *n* − 1, but we used *n* deviations (x – $\overline{x}$) to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other *n* – 1 deviations. The other *n* – 1 deviations can change or vary freely. We call the number *n* – 1 the degrees of freedom.

For example, if we have a sample of size *n* = 20 items, then we calculate the degrees of freedom as *df* = *n* – 1 = 20 – 1 = 19, and we write the distribution as $T \sim t_{19}$.

The following image shows what happens to the t-distribution as you change the degrees of freedom. What happens as the df increases? What happens once n reaches around 30, and how does that relate to what you already know about the CLT?

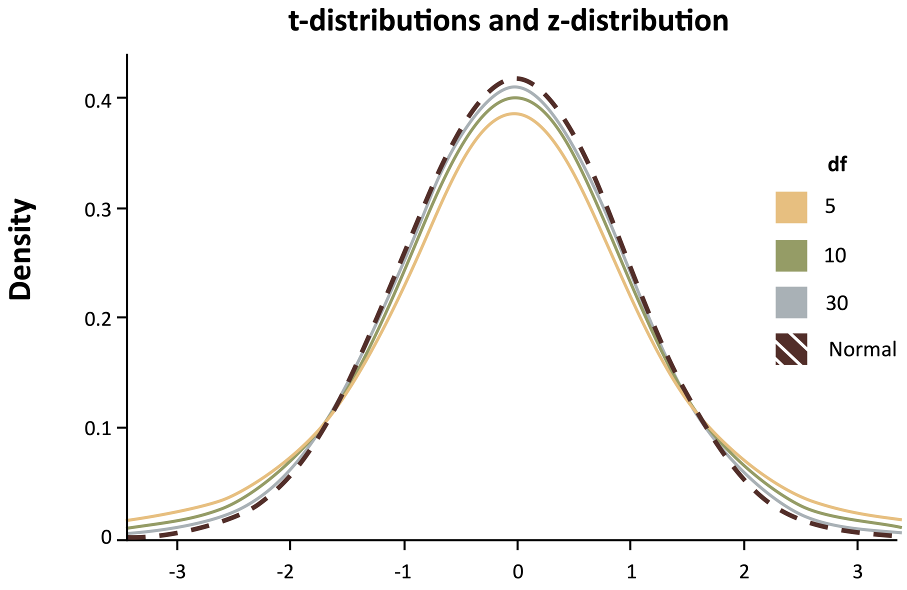


Figure 6.3: t-distribution with different degrees of freedom. *[Figure description available at the end of the section](#)*.

Properties of the Student's *t*-Distribution

To summarize the properties of the t-distribution:

- The graph for the Student's t-distribution is similar to the standard normal curve, in that it is symmetric about a mean of zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell-shaped (normal), then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.
- The notation for the Student's t-distribution (using T as the random variable) is $T \sim t_{df}$ where $df = n - 1$.

Example

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Plots of the data show no skewness or outliers. Which distribution is appropriate to use here?

Solution
You should use the *t*-distribution with *df* = 14 since we do not have information about the population, specifically the standard deviation, and have a small sample (*n* = 15).

Your Turn!

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects and plots of the data show no skewness or outliers. Which distribution is appropriate to use here?

Finding *t*-Distribution Probabilities

A probability table for the Student's *t*-distribution can also be used. The table gives *t*-scores that correspond to the confidence level (column) and degrees of freedom (row). When using a *t*-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails. Notice that most *t*-tables gives *t*-scores given the degrees of freedom and the right-tailed probability.

You'll find that *t*-tables are adequate for finding critical values but are very limited when trying to find ***p*-values**. Calculators and computers can easily calculate any Student's *t*-probabilities.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 6.1: Phillip Glickman (2019). *clear glass cup close-up photography*. Unsplash license. https://unsplash.com/photos/4wnZbnW9Bv0

Figure 6.2: Kindred Grey (2021). *Comparing the standard normal distribution and Student's t-distribution.* CC BY-SA 4.0.

Figure 6.3: Kindred Grey (2021). *t-distribution with different degrees of freedom.* CC BY-SA 4.0.

Figure Descriptions

Figure 6.1: A Guinness Draught beer in a glass next to a candle in an English pub.

Figure 6.2: Two lines on one x, y plot that both follow the bell curve. X axis ranges from negative three to positive three by one. Density is on the Y axis and goes from zero to 0.4 by .1. Normal distribution is taller at the maximum and more narrow on the sides. t distribution is shorter at the maximum point and wider on both sides.

Figure 6.3: Four lines on one x, y plot. X axis ranges from negative three to positive three by one. Density is on the Y axis and goes from zero to 0.4 by .1. All four lines follow bell curve and are very similar. From most density at the maximum point to lowest: Normal, df = 30, df = 10, df = 5.

6.2 Inference for the Mean in Practice

We have discussed how the sampling distribution of the sample mean follows a normal distribution when the population standard deviation, σ, is known and the t-distribution when it is not. In practice, we rarely know the population standard deviation. For larger samples, we can typically get away with using z according to the CLT. In summary, the majority of the time in which we opt to use t, we do not know σ and we have a small sample (n < 30).

Confidence Intervals for the Mean (σ Unknown)

The general format of a **confidence interval** is:

$$PE - MoE, PE + MoE$$

The population parameter is μ. The **point estimate** (PE) for μ is $\overline{x}$, the sample mean.

If the population standard deviation is not known, the **margin of error** for a population mean is:

$$\text{MoE} = (t_{\frac{\alpha}{2}})\,\frac{s}{\sqrt{n}})$$

- $t_{\frac{\alpha}{2}}$ is the t critical value with area to the right equal to $\frac{\alpha}{2}$
- use $df = n - 1$ degrees of freedom
- s = sample standard deviation

Example

The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.[1]

The FEC has reported financial information for 556 Leadership PACs that operated during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 30 Leadership PACs (in dollars).

Receipt data				
\$46,500.00	\$0	\$40,966.50	\$105,887.20	\$5,175.00
\$29,050.00	\$19,500.00	\$181,557.20	\$31,500.00	\$149,970.80
\$2,555,363.20	\$12,025.00	\$409,000.00	\$60,521.70	\$18,000.00
\$61,810.20	\$76,530.80	\$119,459.20	\$0	\$63,520.00
\$6,500.00	\$502,578.00	\$705,061.10	\$708,258.90	\$135,810.00
\$2,000.00	\$2,000.00	\$0	\$1,287,933.80	\$219,148.30

Figure 6.4: PAC receipt data

$\overline{x}$ = \$251,854.23

s = \$521,130.41

Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's *t*-distribution.

Note that we are not given the population standard deviation, only the standard deviation of the sample.

Solution

There are 30 measures in the sample, so n = 30, and df = 30 – 1 = 29.

CL = 0.96, so α = 1 – CL = 1 – 0.96 = 0.04.

$\alpha/2 = 0.02$ $t_{\alpha/2} = t_{0.02} = 2.150$

Margin of error = $t_{\alpha/2}(\frac{s}{\sqrt{n}})$ = 2.150 (521130.41 ÷ √30) $\sim$ \$204,561.66

$\overline{x}$ – margin of error = 251,854.23 – 204,561.66 = \$47,292.57

$\overline{x}$ + margin of error = 251,854.23 + 204,561.66 = \$456,415.89

We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between 47,292.57 and 456,415.89 dollars.

The 96% confidence interval is (\$47,262, \$456,447).

The difference between solutions arises from rounding differences.

Your Turn!

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in figure 6.5. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

TV data				
0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Figure 6.5: Student TV data

Hypothesis Tests for the Mean (σ Unknown)

Remember, we will use the *t*-distribution when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.

If we are testing a single population mean, and we decide to use *t*, the steps say the same, but our test statistic will change slightly.

$$t = \frac{\overline{x}-\mu_o}{\left(\frac{s}{\sqrt{n}}\right)}$$

You should have no problem using technology to find *p*-values associated with a *t*-test statistic. However, if you want to use your *t*-table, you'll find it is somewhat limited in finding exact *p*-values. Despite that, you can still estimate a range of values for your *p*-value and then compare the range to your significance level.

Examples

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores below:

65, 65, 70, 67, 66, 63, 63, 68, 72, 71

Perform the hypothesis test using a 5% level of significance to test the instructor's claim.

Solution
Set up the hypothesis test:

A 5% level of significance means that α = 0.05. This is a test of a single population mean.

H_0: μ = 65

H_a: μ > 65

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

If you read the problem carefully, you will notice that there is no population standard deviation given. You are only given n = 10 sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where n = 10 and df = 10 − 1 = 9.

Calculate the p-value using the Student's t-distribution:

p-value = $P(\overline{x} > 67)$ = 0.0396 where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpret the p-value:

If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.

Compare α and the *p*-value:

Since α = 0.05 and p-value = 0.0396, α > p-value.

Make a decision:

Since α > p-value, reject H_0.

This means you reject μ = 65. In other words, you believe the average test score is actually more than 65.

Conclusion:

At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

Your Turn!

It is believed that a stock price for a particular company will grow at a rate of $5 per week. An investor believes the stock won't grow as quickly. Changes in the stock price are recorded for ten weeks and are as follows:

$4, $3, $2, $3, $1, $7, $2, $1, $1, $2

Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the type I and type II errors.

Summary of Assumptions

When you perform inference on a single population mean μ using a Student's *t*-distribution (often called a *t*-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample standard deviation to approximate the population standard deviation. (Note that, if the sample size is sufficiently large, a *t*-test will work even if the population is not approximately normally distributed).

When you perform a hypothesis test of a single population mean μ using a normal distribution (often called a *z*-test), you take a simple random sample from the population. The population you are testing is normally distributed, or your sample size is sufficiently large. You know the value of the population standard deviation, which is rarely known in reality.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Notes

[1] "Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2, 2013).

6.3 The Sampling Distribution of the Sample Proportion

We have now talked at length about the basics of inference on the mean of quantitative data. What if the variable of interest is categorical? We cannot calculate means, variances, and the like for categorical data. However, we can count the number of individuals that have a certain characteristic and divide by the total number in our population to get the **population proportion (*p*)**.

Understanding the Variability of a Proportion

Suppose we know that the proportion of American adults who support the expansion of solar energy is p = 0.88, which is our parameter of interest. If we were to take a poll of 1,000 American adults on this topic, the estimate would not be perfect, but how close to 88% might we expect the sample proportion to be? We want to understand how the sample proportion, $\hat{p}$, behaves when the true population proportion is 0.88. We can simulate responses we would get from a simple random sample of 1,000 American adults, which is only possible because we know the actual support for expanding solar energy is 0.88. Here's how we might go about constructing such a simulation:

1. There were about 250 million American adults in 2018. On 250 million pieces of paper, write "support" on 88% of them and "not" on the other 12%.
2. Mix up the pieces of paper and pull out 1,000 pieces to represent our sample of 1,000 American adults.
3. Compute the fraction of the sample that say "support."

Any volunteers to conduct this simulation? Probably not. Running this simulation with 250 million pieces of paper would be time-consuming and very costly, but we can simulate it using technology. In this simulation, one sample gave a point estimate of $\hat{p}_1$ = 0.894. We know the population proportion for the simulation was p = 0.88, so we know the estimate had an error of 0.894 − 0.88 = +0.014. One simulation isn't enough to get a great sense of the distribution of estimates we might expect in the simulation, so we should run more simulations. In a second simulation, we get $\hat{p}_2$ = 0.885, which has an error of +0.005. In another, $\hat{p}_3$ = 0.878 gives an error of -0.002. And in another, an estimate of $\hat{p}_4$ = 0.859 means an error of -0.021. With the help of a computer, we've run the simulation 10,000 times and created a histogram of the results from all 10,000 simulations in the figure 6.6.

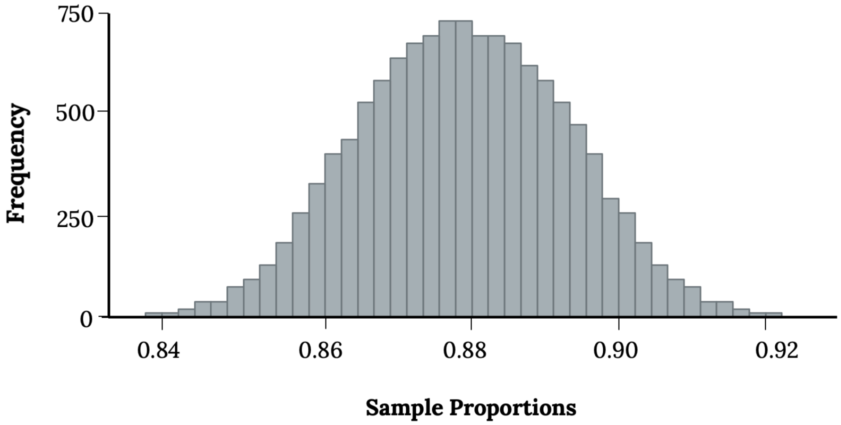


Figure 6.6: Histogram from simulation. [Figure description available at the end of the section](#).

This simulates the sampling distribution of the sample proportion. We can characterize this sampling distribution as follows:

- **Center:** The center of the distribution is $\overline{x}_{\hat{p}}$ = 0.880, which is the same as the parameter. Notice that the simulation mimicked a simple random sample of the population, which is a straightforward sampling strategy that helps avoid sampling bias.
- **Spread:** The standard deviation of the distribution is $s_{\hat{p}}$ = 0.010. When we're talking about a sampling distribution or the variability of a point estimate, we typically use the term "standard error" rather than "standard deviation," and the notation $SE_{\hat{p}}$ is used for the standard error associated with the sample proportion.
- **Shape:** The distribution is symmetric and bell-shaped, and it resembles a normal distribution.

When the population proportion is p = 0.88 and the sample size is n = 1,000, the sample proportion $\hat{p}$ looks to give an unbiased estimate of the population proportion and resembles a normal distribution. It looks as if we can apply the **central limit theorem** here too under the conditions discussed in the following section of this chapter.

Conditions for the CLT for *p*

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with parameters:

$$\mu_{\hat{p}} = \text{p}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{\text{p(1-p)}}{n}}$$

In order for the central limit theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1 - p) \geq 10$. Note that some resources may use 5, but 10 is safer. Hopefully, you see some similarity to the normal approximation to the binomial, which is the underlying idea. It is also typically recommended that the number of successes (x) and failures (n-x) both exceed 10 as well, resulting in a minimum sample size of 20 because the normal approximation just doesn't work well with smaller sample sizes.

What if we do not meet these conditions? Consider the following distributions, and see if any patterns emerge.

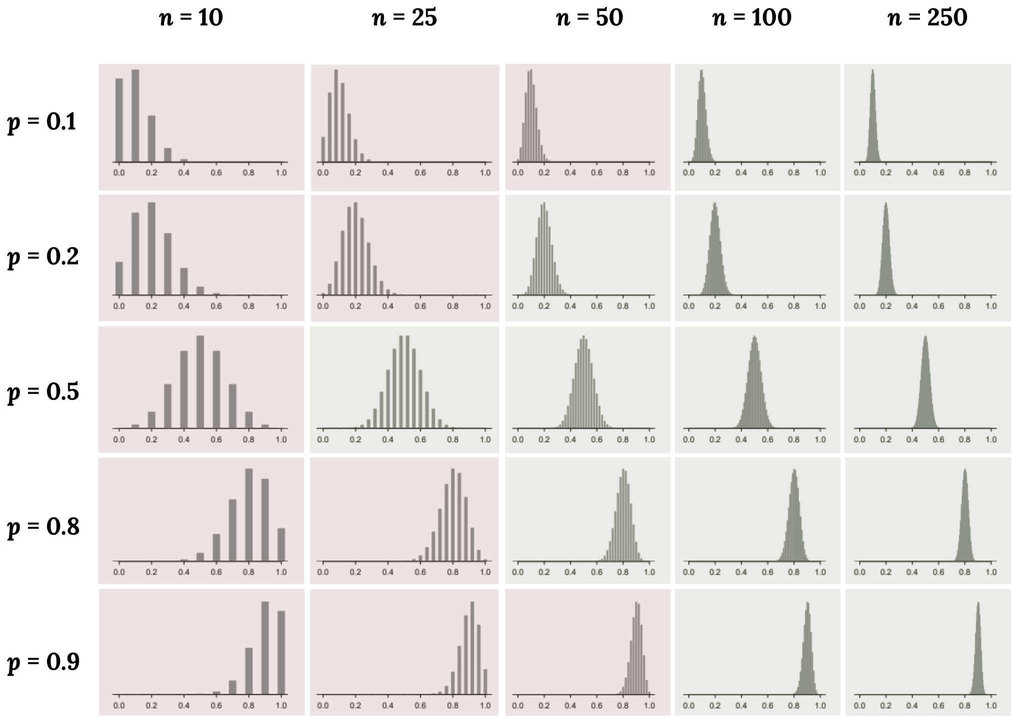


Figure 6.7: Sample size conditions. Figure description available at the end of the section.

The figures above shaded in red do not meet conditions, while the ones in green do. From these distributions, we can see some patterns:

1. When either n is small resulting in np $n(1 - p)$ also being small, the distribution looks more discrete (i.e., not continuous).
2. When np or $n(1 - p)$ is smaller than 10, the skew in the distribution is more noteworthy.
3. The larger both np and $n(1 - p)$, the more normal the distribution. This may be a little harder to see for the larger sample size in these plots as the variability also becomes much smaller.
4. When np and $n(1 - p)$ are both very large, the distribution's discreteness is hardly evident, and the distribution looks much more like a normal distribution.

In regards to how the mean and standard error of the distributions change:

1. The centers of the distribution are always at the population proportion, p, that was used to generate the simulation. Because the sampling distribution of $\hat{p}$ is always centered at the population parameter, p, it means the sample proportion ($\hat{p}$) is accurate (unbiased) when the data are independent and drawn from such a population.
2. For a particular population proportion, the variability in the sampling distribution decreases as the sample size becomes larger. This will likely align with your intuition that an estimate based on a larger sample size will tend to be more accurate.
3. For a particular sample size, the variability will be largest when $p = 0.5$. The differences may be a little subtle, so take a close look. This reflects the role of the proportion p in the standard error formula. The standard error is largest when $p = 0.5$.

At no point will the distribution of $\hat{p}$ look perfectly normal, since $\hat{p}$ will always be take discrete values (x/n). It is always a matter of degree, and we will use the standard success-failure condition with minimums of 10 for np and $n(1 - p)$ as our guideline within this book.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 6.6: Kindred Grey (2020). *Histogram from simulation.* CC BY-SA 4.0.

Figure 6.7: Kindred Grey (2024). *Sample size conditions.* CC BY-SA 4.0. Adaptation of Figures 5.4 and 5.5 from OpenIntro Introductory Statistics (2019) (CC BY-SA 3.0). Retrieved from https://www.openintro.org/book/os

Figure Descriptions

Figure 6.6: Bar chart with narrow bars that follow the normal distribution. The x axis is labeled 'sample proportions' and ranges from 0.84 to 0.92 by .02. The y axis is labeled 'frequency' and ranges from zero to 750 by 250.

Figure 6.7: 20 different bar charts showing that as sample size increases, the bell curve shape and narrowness of the curve increases. These graphs also show that as p increases, the graph shifts from right skewed (p = 0.1) to normal (p = 0.5) to left skewed (p = 0.9).

6.4 Inference for a Proportion

If we are working with **categorical data** our parameter of interest is often the **population proportion**, p. The point estimate for p is $\hat{p} = \frac{x}{n}$, where x is the number of successes and n is the sample size. It is also sometimes denoted as p'. We saw previously that, if we meet conditions $np \geq 10$ and $n(1 - p) \geq 10$, we can apply the **central limit theorem** and assume:

$$\hat{p} \sim \mathrm{N}\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

How do you know you are dealing with a proportion problem? First, the underlying distribution is a **binomial distribution**. This will be categorical data with no mention of a mean or average. If X is a binomial random variable, then X ~ $B(n, p)$, where n is the number of trials and p is the probability of a success.

Hypothesis Tests for *p*

When you perform a **hypothesis test** of a single population proportion p, the steps are exactly the same as what we have seen before; however, we will calculate our **test statistic** differently. When conducting a test for p, our hypotheses will look as follows:

- H_o: $p = p_0$
- H_a: p $(<,>,\neq)$ p_0

Recall, the general form of a test statistic is:

$$Z = \frac{\text{point estimate - null value}}{\text{SE}}$$

For the normal distribution of proportions and if

$$\hat{p} \sim \mathrm{N}\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

then the z-score formula is:

$$\mathrm{z} = \frac{\hat{p}\text{-p}}{\sqrt{\frac{pq}{n}}}$$

Intuitively, you might think we use this as our test statistic, but remember two things:

1. We do not actually know p.
2. In a hypothesis test, we begin by assuming the null is true.

In keeping with these facts, we substitute in p_0 for p in the standard error, which gives us:

$$\sigma_{\hat{p}} = \sqrt{\frac{p_o(1\text{-}p_o)}{n}}$$

We can then find a p-value and make our decision as normal.

Example

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine whether the percentage is 50%. Joon samples 100 first-time brides, and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Solution
Set up the hypothesis test:

The 1% level of significance means that α = 0.01. This is a test of a single population proportion.

H_0: $p = 0.50$

H_a: $p \neq 0.50$

The words "is the same or different from" tell you this is a two-tailed test.

Distribution for the test:

The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P′, the estimated proportion.

$\hat{p} \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$

Therefore, $\hat{p} \sim N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}}\right)$ where $p = 0.50$, $q = 1 - p = 0.50$, $n = 100$, and SE = 0.05.

Calculate the *p*-value using the normal distribution for proportions:

$x = 53$

$\hat{p} = \frac{x}{n} = \frac{53}{100} = 0.53.$

$Z = \frac{0.53 - 0.5}{0.05} = 0.6$

p-value = 2*P ($\hat{p}$ > 0.53) = 0.5485

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value = 0.5485, $\alpha < p$-value.

Make a decision:

Since $\alpha < p$-value, you cannot reject H_0.

Conclusion:

At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated using technology.

Your Turn!

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. Two hundred American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p-value, state your conclusion, and identify the type I and type II errors.

Confidence Intervals for *p*

During election years, we see newspaper articles that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 − 0.03, 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that rise and fall each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that rise and fall each week and for the true proportion of households in the United States that own personal computers.

Constructing Confidence Intervals for *p*

The structure of and procedure to find the confidence interval for a proportion is similar to that for the population mean, but the formulas are different.

The general format of a confidence interval is:

$$\text{PE} - \text{MoE}, \text{PE} + \text{MoE}$$

The population parameter is $\hat{p}$. The point estimate for p is $\hat{p}$, the sample proportion.

The margin of error for a proportion is:

$$\text{MoE} = (z_{\frac{\alpha}{2}}(\sqrt{\frac{\hat{p}\hat{q}}{n}}), \text{ where } \hat{q} = 1 - \hat{p}$$

This formula is similar to the margin of error formula for a mean, except that the "appropriate standard error" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

However, in the margin of error formula, we use $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ as the standard deviation instead of $\sqrt{\frac{pq}{n}}$.

In the margin of error formula, the sample proportions $\hat{p}$ and $\hat{q}$ are estimates of the unknown population proportions p and q. The estimated proportions $\hat{p}$ and $\hat{q}$ are used because p and q are not known. The sample proportions $\hat{p}$ and $\hat{q}$ are calculated from the data; $\hat{p}$ is the estimated proportion of successes, and $\hat{q}$ is the estimated proportion of failures.

Example

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 respond that they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution

Let X = the number of people in the sample who have cell phones. X is binomial. X~$B(500, \frac{421}{500})$.

To calculate the confidence interval, you must find p', q', and EBP.

$n = 500$

x = the number of successes = 421

$\hat{p} = \frac{x}{n} = \frac{421}{500} = 0.842$

$\hat{p} = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$\hat{q} = 1 - \hat{p} = 1 - 0.842 = 0.158$

Since CL = 0.95, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$ $(\frac{\alpha}{2}) = 0.025$.

Then $z_{\alpha/2} = z_{0.025} = 1.96$

$$\text{EBP} = (z_{\alpha/2})\sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$\hat{p}$ − EBP = 0.842−0.032 = 0.81

$\hat{p}$ + EBP = 0.842+0.032 = 0.874

The confidence interval for the true binomial population proportion is ($\hat{p}$ − EBP, $\hat{p}$ + EBP) = (0.810, 0.874).

Interpretation:

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% confidence level:

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Your Turn!

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 report owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

6.5 Behavior of Confidence Intervals for a Proportion

Confidence intervals (CI) for p behave similarly to intervals for μ, though there are a few subtle differences.

Calculating the Sample Size

If researchers desire a specific margin of error, then they can use the margin of error formula to calculate the required sample size, n.

Recall the **margin of error** for a population proportion is:

$$\text{MoE} = z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

Solving for n gives you an equation for the sample size:

$$\text{n} = \left(\frac{z_{\frac{\alpha}{2}}}{MoE}\right)^2 p(1 - p)$$

Recall the objective of a CI. If we are looking to estimate p, then we do not know what it is, even though it appears in this formula. So what do we plug in for p? We have a few options:

- If you have prior information, such as a previous sample, and can calculate a point estimate $\hat{p}$, plug it in!
- You can use your best guess at p.
- You can use a "conservative" estimate of p, 0.5.

NOTE: Remember that $\hat{q}$ = (1 – $\hat{p}$), though we do not know $\hat{p}$ yet. Since we multiply $\hat{p}$ and $\hat{q}$ together, we make them both equal to 0.5. Why? Because $\hat{p}$ $\hat{q}$ = (0.5)(0.5) = 0.25 results in the largest possible product. (Try other products: (0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16; and so on.) The largest possible product gives us the largest n. This gives us a large enough sample to be CL% confident that we are within three percentage points of the true population proportion.

Example

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones?

Solution

From the problem, we know that the margin of error is 0.03 (3% = 0.03) and $z_{\alpha/2}$ = $z_{0.05}$ = 1.645 because the confidence level is 90%.

n = $\left(\frac{z}{MoE}\right)^2 \hat{p}\ \hat{q}$ gives n = $\left(\frac{1.645}{0.03}\right)^2$ (0.5)(0.5)

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Your Turn!

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

“Plus Four” Confidence Interval for *p*

This is an alternative *optional* method for constructing a CI for p stemming from the continuity correction of the binomial approximation.

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes, and two are failures. The new sample size is then $n + 4$, and the new count of successes is $x + 2$.

Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example

A random sample of 25 statistics students was asked: “Have you smoked a cigarette in the past week?” Six students reported smoking within the past week. Use the “plus-four” method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution

Six students out of 25 reported smoking within the past week, so $x = 6$ and $n = 25$. Because we are using the “plus-four” method, we will use $x = 6 + 2 = 8$ and $n = 25 + 4 = 29$.

$\hat{p} = \frac{x}{n} = \frac{8}{29} \approx 0.276$

$\hat{q} = 1 - \hat{p} = 1 - 0.276 = 0.724$

Since CL = 0.95, we know $\alpha = 1 - 0.95 = 0.05$ and $\alpha/2 = 0.025$.

$z_{0.025} = 1.96$

Margin of error = $\left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = (1.96)\sqrt{\frac{(0.276)(0.724)}{29}} \approx 0.163.$

$\hat{p}$ − MoE = 0.276 − 0.163 = 0.113

$\hat{p}$ + MoE = 0.276 + 0.163 = 0.439

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Your Turn!

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the “plus-four” method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Chapter 6 Wrap-Up

Concept Check

Take this quiz to check your comprehension of this chapter.

If you are using an offline version of this text, access the quiz for this chapter via the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=320#h5p-162

Section Resources

If you are using an offline version of this text, access these materials by visiting https://doi.org/10.7294/26207456.

6.1 The Sampling Distribution of the Sample Mean (*t*)

6.2 Inference for the Mean in Practice

6.3 The Sampling Distribution of the Sample Proportion

6.4 Inference for a Proportion

6.5 Behavior of Confidence Intervals for a Proportion

Key Terms

Try to define the terms below on your own. Check your response by clicking on the term, or looking at the end-of-book glossary!

6.1 Sampling Distribution for the Sample Mean

- **Sampling distribution**
- **Student's *t*-distribution**
- **Degrees of freedom (df)**
- ***p*-value**

6.2 Inference for the Mean in Practice

- **Confidence interval**
- **Point estimate**
- **Margin of error (MoE)**

6.3 Sampling Distribution of the Sample Proportion

- **Population proportion** (p)
- **Sample proportion** $(\hat{p})$
- **Central limit theorem (CLT)**

6.4 Inference for a Proportion

- **Categorical data**
- **Binomial distribution**
- **Test statistic**

6.5 Behavior of Confidence Intervals for a Proportion

Extra Practice

Extra practice problems are available at the end of the book (Chapter 6 Extra Practice).

CHAPTER 7: INFERENCE FOR TWO SAMPLES

7.1 Inference for Two Dependent Samples (Matched Pairs)

Learning Objectives

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

- Classify hypothesis tests by type
- Conduct and interpret hypothesis tests for two population means with known population standard deviations
- Conduct and interpret hypothesis tests for two population means with unknown population standard deviations
- Conduct and interpret hypothesis tests for matched or paired samples
- Conduct and interpret hypothesis tests for two population proportions

Figure 7.1: Types of breakfasts. If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River), we will use a two-sample analysis. *Figure description available at the end of the section.*

Studies often compare two groups. For example, maybe researchers are interested in the effect aspirin has in preventing heart attacks. One group is given aspirin and the other a **placebo** that has no effect, and the groups' heart attack rates are studied over several years. Other studies may compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

You have learned to conduct **inference** on single means and single proportions. We know that the first step is deciding the type of data with which we are working. For **quantitative data**, we are focused on means, while for **categorical data**, we are focused on proportions. In this chapter, we will compare two means or two proportions. The general procedure is still the same, just expanded. With two-sample analysis, it is good to know what the formulas look like and where they originate; however, you will probably lean heavily on technology in preforming the calculations.

In comparing two means, we are obviously working with two groups, but first we need to think about the relationship between them. The groups are classified either as **independent** or **dependent**. Independent samples consist of two samples that have no relationship—that is, sample values selected from one population are not related in any way to sample values selected from the other population. Dependent samples consist of two groups that have some sort of identifiable relationship.

Two Dependent Samples (Matched Pairs)

Two samples that are dependent typically come from a **matched pairs** experimental design. The parameter tested using matched pairs is the **population mean difference**. When using inference techniques for matched or paired samples, the following characteristics should be present:

- Simple random sampling is used.
- Sample sizes are often small.
- Two measurements (samples) are drawn from the same (or extremely similar) two individuals or objects.
- Differences are calculated from the matched or paired samples.
- The differences form the sample that is used for analysis.

To perform statistical inference techniques, we first need to know about the **sampling distribution** of our parameter of interest. Remember that, although we start with two samples, the differences are the data in which we are interested, and our parameter of interest is μ_d, the mean difference. Our **point estimate** is $\overline{x}_d$. In a perfect world, we could assume that both samples come from a normal distribution; therefore, the differences in those normal distributions are also normal. However, in order to use Z, we must know the population standard deviation, which is near impossible for a difference distribution. In addition, it is very hard to find large numbers of matched pairs, so the sampling distribution we typically use for $\overline{x}_d$ is a *t*-distribution with $n - 1$ degrees of freedom, where n is the number of differences.

Confidence intervals may be calculated on their own for two samples, but often, we first want to conduct a hypothesis test to formally check if a difference exists, especially in the case of matched pairs. If we do find a statistically significant difference, then we may estimate it with a CI after the fact.

Hypothesis Tests for the Mean Difference

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated, and the population mean difference, μ_d, is our parameter of interest. Although it is possible to test for a certain magnitude of effect, we are most often just looking for a general effect. Our hypothesis would then look like:

- H_o: $\mu_d = 0$
- H_a: μ_d (<, >, ≠) 0

The steps are familiar to us by now, but it is tested using a Student's t-test for a single population mean with n – 1 degrees of freedom, with the test statistic:

$$t = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the figure below. A lower score indicates less pain. The "before" value is matched to an "after" value, and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Subject	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Figure 7.2: Reported pain data

Solution

Corresponding "before" and "after" values form matched pairs. (Calculate "after" – "before.")

After data	Before data	Difference
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

Figure 7.3: Differences

The data for the test are the differences: {0.2, −4.1, −1.6, −1.8, −3.2, −2, −2.9, −9.6}.

The sample mean and sample standard deviation of the differences are: $\overline{x}_d$ = −3.13 and s_d = 2.91. Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote "differences."

Random variable: $\overline{x}_d$ = the mean difference of the sensory measurements.

H_0: $\mu_d \geq 0$

The null hypothesis is zero or positive, meaning that there is the same or more pain felt after hypnotism. That means the subject shows no improvement. μ_d is the population mean of the differences.

H_a: $\mu_d < 0$

The alternative hypothesis is negative, meaning there is less pain felt after hypnotism. That means the subject shows improvement. The score should be lower after hypnotism, so the difference ought to be negative to indicate improvement.

Distribution for the test:

The distribution is a Student's *t* with $df = n - 1 = 8 - 1 = 7$. Use t_7. (Notice that the test is for a single population mean.)

Calculate the *p*-value using the Student's-t distribution: *p*-value = 0.0095

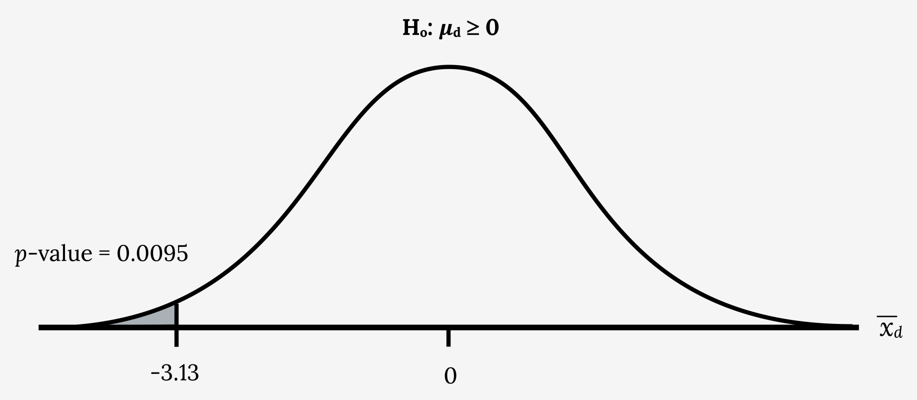


Figure 7.4: Reported pain p-value. Figure description available at the end of the section.

$\overline{x}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$\overline{x}_d = -3.13$

$\overline{s}_d = 2.91$

Compare *α* and the *p*-value:

$\alpha = 0.05$ and *p*-value = 0.0095, so $\alpha >$ *p*-value.

Make a decision:

Since $\alpha >$ *p*-value, reject H_0. This means that $\mu_d < 0$ and there is improvement.

Conclusion:

At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

Note: For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (after – before) and put the differences into a list or you can put the after data into a first list and the before data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name – 2nd list name. The calculator will do the subtraction, and you will have the differences in the third list. Use your list of differences as the data.

Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter `0` for μ0, the name of the list where you put the data, and `1` for Freq:. Arrow down to `μ`: and arrow over to < μ0. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The *p*-value is 0.0094, and the test statistic is -3.04. Do these instructions again except, arrow to `Draw` (instead of `Calculate`). Press `ENTER`.

Your Turn!

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5% level.

Subject	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Figure 7.5: Cholesterol levels

Confidence Intervals for the Mean Difference

The general format of a confidence interval is (PE − MoE, PE + MoE)

The population parameter of interest is μ_d, the mean difference. Our point estimate is $\overline{x}_d$.

If we are using the *t*-distribution, the margin of error for the population mean difference is:

$$\text{MoE} = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s_d}{\sqrt{n}}\right)$$

- $t_{\frac{\alpha}{2}}$ is the t critical value with area to the right equal to $\frac{\alpha}{2}$
- use $df = n - 1$ degrees of freedom, where n is the number of pairs
- s_d = standard deviation of the differences

Example

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	Player 1	Player 2	Player 3	Player 4
Amount of weight lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

Figure 7.6: Weight lifted

The coach wants to know if the strength development class makes his players stronger on average.

Solution
Record the differences data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. Assume the differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation

$\overline{x}_d$ = 21.3, s_d = 46.7

NOTE: The data given here would indicate that the distribution is actually right-skewed. The difference 90 may be an extreme outlier? It is pulling the sample mean to be 21.3 (positive). The means of the other three data values are actually negative.

Using the difference data, this becomes a test of a single variable.

Define the random variable:

$\overline{X}_d$ mean difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

H_0: $\mu_d \leq 0$

H_a: $\mu_d > 0$

Graph:

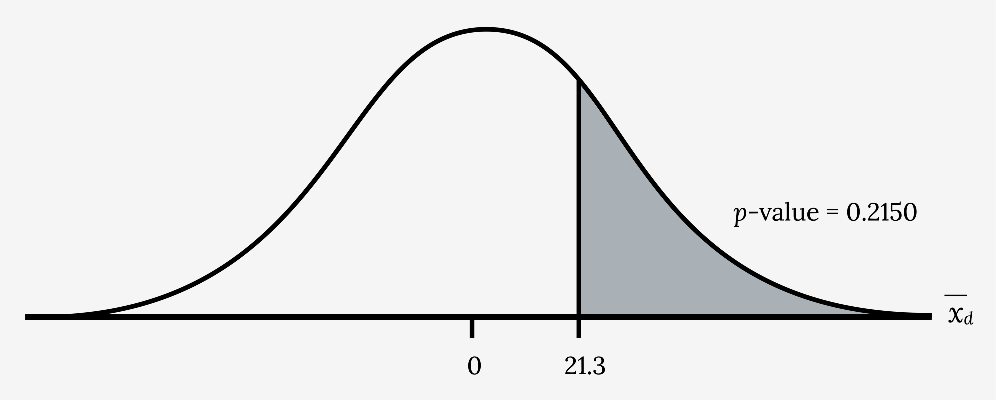


Figure 7.7: Weight lifted p-value. Figure description available at the end of the section.

Calculate the *p*-value.

The p-value is 0.2150.

Decision:

If the level of significance is 5%, the decision is not to reject the null hypothesis, because $\alpha < p$-value.

What is the conclusion?

At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

Your Turn!

A new prep class was designed to improve SAT test scores. Five students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data are recorded in the figure below. Are the scores, on average, higher after the class? Test at a 5% level.

SAT scores	Student 1	Student 2	Student 3	Student 4
Score before class	1840	1960	1920	2150
Score after class	1920	2160	2200	2100

Figure 7.7: SAT scores

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 7.1: Ali Inay (2015). *variety of foods on top of gray table.* Unsplash license. https://unsplash.com/photos/y3aP9oo9Pjc

Figure 7.4: Kindred Grey (2020). *Reported pain p-value.* CC BY-SA 4.0.

Figure 7.7: Kindred Grey (2020). *Weight lifted p-value.* CC BY-SA 4.0.

Figure Descriptions

Figure 7.1: Ariel picture of a table full of breakfast food including waffles, fruit, breads, coffee, etc.

Figure 7.4: Normal distribution curve showing the values zero and -3.13. -3.13 is associated with p-value 0.0095 and everything to the left of this is shaded.

Figure 7.7: Normal distribution curve with values of zero and 21.3. A vertical upward line extends from 21.3 to the curve and the p-value is indicated in the area to the right of this value.

7.2 Inference for Two Independent Sample Means

Suppose we have two **independent** samples of quantitative data. If there is no apparent relationship between the means, our parameter of interest is the difference in means, μ1-μ2, with a point estimate of $\overline{X}_1$ - $\overline{X}_2$.

The comparison of two population means is very common. A difference between the two samples depends on both the means and their respective standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means and divide by the **standard error**, standardizing the difference. We know that when conducting an **inference** for means, the sampling distribution we use (Z or *t*) depends on our knowledge of the population standard deviation.

Both Population Standard Deviations Known (*Z*)

Even though this situation is unlikely since population standard deviations are rarely known, we will begin demonstrating these ideas under the ideal circumstances. If we know both means' **sampling distributions** are normal, the sampling distribution for the difference between the means is normal, and both populations must be normal. We can combine the standard errors of each sampling distribution to get a standard error of:

$$\sqrt{\frac{(\sigma_1)^2}{n_1}+\frac{(\sigma_2)^2}{n_2}}$$

So the sampling distribution of $\overline{X}_1$ - $\overline{X}_2$, assuming we know both standard deviations, is approximately:

$$\text{N}\left(\mu_1-\mu_2,\sqrt{\frac{(\sigma_1)^2}{n_1}+\frac{(\sigma_2)^2}{n_2}}\right)$$

Therefore, the *z*-test statistic would be:

$$z = \frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1}+\frac{(\sigma_2)^2}{n_2}}}$$

Our confidence interval would be in the form (PE − MoE, PE + MoE), where our **point estimate** is $\overline{X}_1$ - $\overline{X}_2$, and the margin of error is made up of:

$$\text{MoE} = \left(z_{\frac{\alpha}{2}}\right)(SE)$$

- $z_{\frac{\alpha}{2}}$ is the z critical value with area to the right equal to $\frac{\alpha}{2}$
- SE is $\sqrt{\frac{(\sigma_1)^2}{n_1}+\frac{(\sigma_2)^2}{n_2}}$

Since we rarely know one population's standard deviation, much less two, the only situation where we might consider using this in practice is for two very large samples.

Both Population Standard Deviations Unknown (*t*)

Most likely, we will not know the population standard deviations, but we can estimate them using the two sample standard deviations from our independent samples. In this case, we will use a *t* sampling distribution with the following standard error:

$$\sqrt{\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}}$$

Assumptions for the Difference in Two Independent Sample Means

Recall that we need to be able to assume an underlying normal distribution and no outliers or skewness in order to use the *t*-distribution. We can relax these assumptions as our sample sizes get bigger and can typically just use the Z distribution for very large sample sizes.

The remaining question concerns what we do for **degrees of freedom** when comparing two groups. One method requires a somewhat complicated calculation, but if you have access to a computer or calculator, this isn't an issue. We can find a precise *df* for two independent samples as follows:

$$\text{df} = \frac{\left(\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2+\left(\frac{1}{n_2-1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

NOTE: The *df* are not always a whole number; you usually want to round down. It is not necessary to compute this by hand. Find a reliable technology to do this.

If you are working on your own without access to technology, the above formula could be daunting. Another method is to use a conservative estimate of the *df*: min(n_1-1, n_2-1).

Hypothesis Tests for the Difference in Two Independent Sample Means

Recall that the steps to a hypothesis test never change. When our parameter of interest is μ_1-μ_2, we are often interested in an effect between the two groups. In order to show an effect, we will have to first assume there is no difference by stating it in the null hypothesis as:

- H_o: $\mu_1 - \mu_2 = 0$ OR H_o: $\mu_1 = \mu_2$
- H_a: $\mu_1 - \mu_2$ (<, >, ≠) 0 OR H_a: μ_1 (<, >, ≠) μ_2

The *t*-test statistic is calculated as follows:

$$\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}}}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively.
- $\overline{x}_1$ and $\overline{x}_2$ are the sample means. μ_1 and μ_2 are the population means. (NOTE: in the null, we are typically assuming $\mu_1 - \mu_2 = 0$.)

Confidence Intervals for the Difference in Two Independent Sample Means

Once we have identified a difference in a hypothesis test, we may want to estimate it. Our **confidence interval** would be of the form (PE − MoE, PE + MoE), where our point estimate is $\overline{x}_1 - \overline{x}_2$, and the MoE is made up of:

$$\text{MoE} = \left(t_{\frac{\alpha}{2}}\right)(SE)$$

- $t_{\frac{\alpha}{2}}$ is the t critical value with area to the right equal to $\frac{\alpha}{2}$
- SE is $\sqrt{\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}}$

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

7.3 Inference for Two-Sample Proportions

Comparing two proportions, like comparing two means, is also very common when we are working with categorical data. If our parameter of inference is $p_1 - p_2$, then we can estimate it with $\hat{p}_1$ - $\hat{p}_2$.

When conducting inference on two independent **population proportions**, the following characteristics should be present:

- The two independent samples are simple random samples that are independent.
- For each of the samples, the number of successes is at least five, and the number of failures is at least five.
- Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Sampling Distribution of the Difference in Two Proportions

We can build a **sampling distribution** for $\hat{p}_1$ - $\hat{p}_2$ similar to what we did for the difference in two independent sample means. The difference of two proportions follows an approximate normal distribution. We will wait to show the standard error and sampling distribution because we calculate them slightly differently for hypothesis tests and confidence intervals.

Hypothesis Test for the Difference in Two Proportions

If two estimated proportions are different, it may be due to a difference in the populations, or it may be due to chance. A **hypothesis test** can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

Generally, the null hypothesis states that the two proportions are the same (i.e., H_0: $p_1 = p_2$). Since we are assuming there is no difference in the null, we can use both samples to estimate the pooled proportion, p_p, calculated as follows:

$$p_\text{p} = \frac{x_1+x_2}{n_1+n_2}$$

We can use this **pooled proportion** in the calculation of our z-test statistic:

$$z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{p_p(1-p_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Example

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given Medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given Medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

Solution
The problem asks for a difference in proportions, making it a test of two proportions.

Let A and B be the subscripts for medication A and medication B, respectively. Then p_A and p_B are the desired population proportions.

Random variable: $\hat{p}_A - \hat{p}_B$ = difference in the proportions of adult patients who did not react after 30 minutes to medication A and to medication B.

H_0: $\hat{p}_A = \hat{p}_B$ or $\hat{p}_A - \hat{p}_B = 0$

H_A: $\hat{p}_A \neq \hat{p}_B$ or $\hat{p}_A - \hat{p}_B \neq 0$

The words "is a difference" tell you the test is two-tailed.

Distribution for the test:

Since this is a test of two binomial population proportions, the distribution is normal.

Find the pooled proportion: p_p

$p_p = \frac{x_A + x_B}{n_A + n_B} = \frac{20+12}{200+200} = 0.08$

$1 - p_p = 0.92$

$\hat{p}_A - \hat{p}_B$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution:

p-value = 0.1404

Estimated proportion for group A: $\hat{p}_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $\hat{p}_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

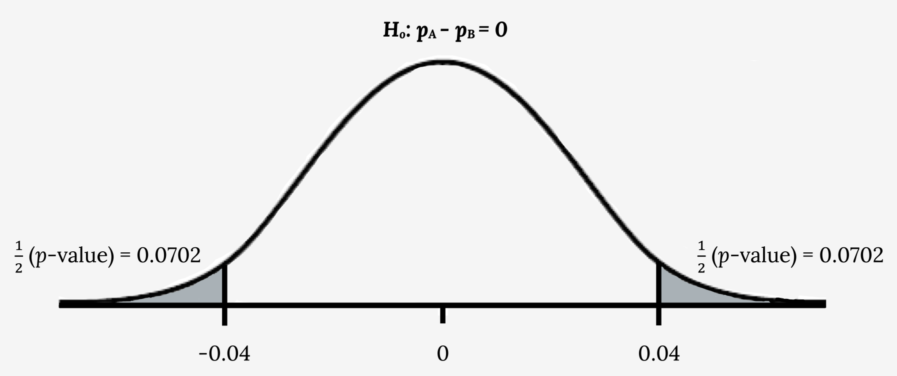


Figure 7.8: Medications A and B. Figure description available at the end of the section.

$\hat{p}_A - \hat{p}_B$ = 0.1 – 0.06 = 0.04.

Half the p-value is below −0.04, and half is above 0.04.

Compare α and the p-value:

α = 0.01 and the p-value = 0.1404. $\alpha < p$-value.

Make a decision:

Since $\alpha < p$-value, do not reject H_0.

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

Your Turn!

Two types of valves are being tested to determine if there is a difference in pressure tolerances. For Valve A, 15 out of a random sample of 100 cracked under 4,500 psi. For Valve B, six out of a random sample of 100 cracked under 4,500 psi. Test at a 5% level of significance.

Confidence Intervals for the Difference in Two Proportions

Once we have identified the presence of a difference in a two-sample test, we may want to estimate it. Our **confidence interval** would be of the form (PE − MoE, PE + MoE), where our point estimate is $\hat{p}_1$ - $\hat{p}_2$, and the MoE is made up of:

$$\text{MoE} = \left(z_{\frac{\alpha}{2}}\right)(SE)$$

- $z_{\frac{\alpha}{2}}$ is the z critical value with area to the right equal to $\frac{\alpha}{2}$
- SE $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - In the SE will we estimate p_1 with $\hat{p}_1$ and p_2 with $\hat{p}_2$ if we do not know them.

Putting that all together, our formula for a CI to estimate the difference in two proportions will be:

$$\hat{p}_1 - \hat{p}_2 \pm \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{\hat{p}_1(1\text{-}\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1\text{-}\hat{p}_2)}{n_2}}$$

Additional Resources

Click here for additional multimedia resources, including podcasts, videos, lecture notes, and worked examples.

If you are using an offline version of this text, access the resources for this section via the QR code, or by visiting https://doi.org/10.7294/26207456.

Figure References

Figure 7.8: Kindred Grey (2020). *Medications A and B*. CC BY-SA 4.0.

Figure Descriptions

Figure 7.8: Normal distribution curve of the difference in the percentages of adult patients who don't react to medication A and B after 30 minutes. The mean is equal to zero, and the values -0.04, 0, and 0.04 are labeled on the horizontal axis. Two vertical lines extend from -0.04 and 0.04 to the curve. The region to the left of -0.04 and the region to the right of 0.04 are each shaded to represent 1/2(p-value) = 0.0702.

Chapter 7 Wrap-Up

Concept Check

Take this quiz to check your comprehension of this chapter.

If you are using an offline version of this text, access the quiz for this chapter via the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=370#h5p-171

Section Resources

If you are using an offline version of this text, access these materials by visiting https://doi.org/10.7294/26207456.

7.1 Inference for Two Dependent Samples (Matched Pairs)

7.2 Inference for Two Independent Sample Means

7.3 Inference for Two-Sample Proportions

Key Terms

Try to define the terms below on your own. Check your response by clicking on the term, or looking at the end-of-book glossary!

7.1 Inference for Two Dependent Samples (Matched Pairs)

- **Placebo**
- **Inference**
- **Quantitative data**
- **Categorical data**
- **Independence**
- **Matched pairs**
- **Population mean difference**
- **Sampling distribution**
- **Point estimate**

7.2 Inference for Two Independent Sample Means

- **Standard error**
- **Degrees of freedom**
- **Confidence interval**

7.3 Inference for Two-Sample Proportions

- **Population proportion**
- **Pooled proportion**

Extra Practice

Extra practice problems are available at the end of the book (Chapter 7 Extra Practice).

EXTRA PRACTICE

Chapter 1 Extra Practice

1.1 Introduction to Statistics

1. Determine how the key terms apply to the following study. We want to know the average (mean) amount of money first-year college students spend at ABC College on school supplies (excluding books). We randomly survey 100 first-year students at the college. Three of those students spent $150, $200, and $225.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=1070#h5p-2

---

2. Determine how the key terms apply to the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95.

---

3. Determine how the key terms apply to the following study.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=1070#h5p-13

4. As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of automobile crashes on test dummies. Here is the criterion they used:

Speed at which cars crashed	Location of "driver" (i.e., dummies)
35 miles/hour	Front seat

Figure 1.12

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries if they had been actual drivers. We start with a simple random sample of 75 cars.[1]

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=1070#h5p-14

---

5. An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines how many in the sample have been involved in malpractice lawsuits.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=1070#h5p-15

6. Pharmaceutical companies often conduct surveys to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months that patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data was collected, measuring the time (in months) between patients starting treatment and their deaths.

Researcher A: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

Researcher B: 3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44, 23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

Determine how the key terms apply to the example for Researcher A.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=1070#h5p-16

7. For each of the following exercises, identify:

a. The population
b. The sample
c. The parameter
d. The statistic
e. The variable
f. The data

Give examples where appropriate.

i. A fitness center is interested in the mean amount of time a client exercises in the center each week.

ii. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

iii. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

iv. Insurance companies are interested in clients' mean yearly health costs so that they can determine the costs of health insurance.

v. A politician is interested in the proportion of voters in his district who think he is doing a good job.

vi. A marriage counselor is interested in the proportion of clients she counsels who stay married.

vii. Political pollsters may be interested in the proportion of people who will vote for a particular cause.

viii. A marketing company is interested in the proportion of people who will buy a particular product.

---

8. A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

i. What is the population in which she is interested?

a. All Lake Tahoe Community College students
b. All Lake Tahoe Community College English students
c. All Lake Tahoe Community College students in her classes
d. All Lake Tahoe Community College math students

ii. Consider the following: X = number of days a Lake Tahoe Community College math student is absent. In this case, X is an example of a:

a. Variable
b. Population
c. Statistic
d. Data

iii. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. Parameter
b. Data
c. Statistic
d. Variable

9. In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9%.

i. The “average increase” for all NASDAQ stocks is the:

a. Population
b. Statistic
c. Parameter
d. Sample
e. Variable

ii. All of the NASDAQ stocks are the:

a. Population
b. Statistic
c. Parameter
d. Sample
e. Variable

iii. Nine percent is the:

a. Population
b. Statistic
c. Parameter
d. Sample
e. Variable

iv. The 100 NASDAQ stocks in the survey are the:

a. Population
b. Statistic
c. Parameter
d. Sample
e. Variable

v. The percent increase for one stock in the survey is the:

a. Population
b. Statistic
c. Parameter
d. Sample
e. Variable

vi. Would the data collected be qualitative, quantitative discrete, or quantitative continuous?

1.2 Data Basics

1. The data are the colors of backpacks. You sample five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. What type of data is this?

2. The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. What type of data are the numbers of books (three, four, two, and one)?

3. The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, and 4.3. Notice that backpacks carrying three books can have different weights. What type of data is this?

4. The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

5. The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

6. The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

7. Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. If they are quantitative data, indicate whether they are continuous or discrete.

1.3 Data Collection and Observational Studies

1. Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups; one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.[2]

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=1070#h5p-17

---

2. A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

---

3. You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

a. Describe the explanatory and response variables in the study.
b. What are the treatments?
c. What should you consider when selecting participants?
d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
e. Identify any lurking variables that could interfere with this study.
f. How can blinding be used in this study?

4. Identify any issues with the following studies.

a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
b. A research study is designed to investigate a new children's allergy medication.
c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

1.5 Sampling

1. Determine whether or not the following samples are representative.

a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
b. To determine the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every 20th child under age ten who enters the supermarket.
c. To find the average annual income of all adults in the United States, sample US congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every US congressman in the cluster.
d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

2. Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.lib.vt.edu/significantstatistics/?p=1070#h5p-20

3. A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities. What type of sampling is used—simple random, stratified, systematic, cluster, or convenience?

4. This table displays six sets of quiz scores for an elementary-level statistics class, with each quiz worth a possible ten points. Use the random number generator to generate different types of samples from the data.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Figure 1.13

a. Create a stratified sample by column. Pick three quiz scores randomly from each column.
b. Create a cluster sample by picking two of the columns. Use the column numbers one through six.
c. Create a simple random sample of 15 quiz scores.
d. Create a systematic sample of 12 quiz scores.

5. Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first-term organic chemistry class. Many of these students are taking first-term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128 $87 $173 $116 $130 $204 $147 $189 $93 $153

The second sample is taken using a list of senior citizens who take P.E. classes, selecting every fifth senior citizen on the list until we have a total of ten senior citizens.

They spend:

$50 $40 $36 $15 $50 $100 $40 $53 $22 $22

It is unlikely that any student is in both samples.

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180 $50 $150 $85 $260 $75 $180 $200 $200 $150

Is the sample biased?

---

6. What type of data is "number of times per week?"

---

7. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in Norfolk, Virginia. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

a. What was the sampling method used?
b. "Duration (amount of time)" is what type of data?
c. The colors of the houses around the park are what kind of data?
d. What is the population?

8. The following figure contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.[3]

Year	Total number of deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	**823,856**

Figure 1.14

Use figure above to answer the following questions.

a. What is the proportion of deaths between 2007 and 2012?
b. What percent of deaths occurred before 2001?
c. What is the percent of deaths that occurred in 2003 or after 2010?
d. What is the fraction of deaths that occurred before 2012?
e. What kind of data is the number of deaths?
f. Earthquakes are quantified according to the amount of energy they produce (e.g., 2.1, 5.0, 6.7). What type of data is that?
g. What contributed to the large number of deaths in 2010? In 2004? Explain.

---

9. Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A group of test subjects is divided into 12 groups; then four of the groups are chosen at random.
b. A market researcher polls every tenth person who walks into a store.
c. The first 50 people who walk into a sporting event are polled on their television preferences.
d. A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

10. Pharmaceutical companies often conduct studies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months that patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data was collected, measuring the time (in months) between patients starting treatment and their deaths.

Researcher A: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

Researcher B: 3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44, 23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

a. List two reasons why the data may differ.
b. Can you tell if either researcher is correct or incorrect? Why?
c. Would you expect the data to be identical? Why or why not?
d. Suggest at least two methods the researchers might use to gather random data.
e. Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?
f. Suppose that the second researcher conducted his survey by choosing 40 patients he personally knew. What sampling method would that researcher have used? What concerns would you have about this dataset based upon the data collection method?

---

11. Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school, collecting the following data.

Hours played per week	Frequency	Relative frequency	Cumulative relative frequency
0-2	26	0.17	0.17
2-4	30	0.20	0.37
4-6	49	0.33	0.70
6-8	25	0.17	0.87
8-10	12	0.08	0.95
10-12	8	0.05	1

Figure 1.15: Researcher A

Hours played per week	Frequency	Relative frequency	Cumulative relative frequency
0-2	48	0.32	0.32
2-4	51	0.34	0.66
4-6	24	0.16	0.82
6-8	12	0.08	0.90
8-10	11	0.07	0.97
10-12	4	0.03	1

Figure 1.16: Researcher B

a. Give a reason why the data may differ.
b. Would the sample size be large enough if the population is the students in the school?
c. Would the sample size be large enough if the population is school-aged children and young adults in the United States?
d. Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?
e. As a way of rewarding students for participating in the survey, the researchers gave each student a gift card to a video game store. Would it affect the data if students knew about the award before the study?

---

12. A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in Figure 1.17. The second study collected the data in Figure 1.18.

Group	Showed improvement	No improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Figure 1.17: First study

Group	Showed improvement	No improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

Figure 1.18: Second study

a. Given what you know, which study is correct?
b. The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?

c. Both groups that performed the study concluded that the software works. Is this accurate?
d. The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?
e. Patients who used the software were also a part of an exercise program, whereas patients who did not use the software were not. Does this change the validity of the conclusions from the second study?
f. Is a sample size of 1,000 a reliable measure for a population of 5,000?
g. Is a sample of 500 volunteers a reliable measure for a population of 2,500?
h. A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?
i. Is a sample size of two representative of a population of five?
j. Is it possible for two well-run experiments with similar sample sizes to get different data?

---

13. For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of that data.

a. Number of tickets sold to a concert
b. Percent of body fat
c. Favorite baseball team
d. Time in line to buy groceries
e. Number of students enrolled at Evergreen Valley College
f. Most watched television show
g. Favorite brand of toothpaste
h. Distance to the closest movie theatre
i. Age of executives in Fortune 500 companies
j. Number of competing computer spreadsheet software packages

---

14. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in Norfolk. The first house in the neighborhood around the park was selected randomly, and then every eighth house in the neighborhood around the park was interviewed.

a. "Number of times per week" is what type of data?
b. "Duration (amount of time)" is what type of data?

---

15. Airline companies are interested in the consistency of the number of babies on each flight so that they have adequate safety equipment. Suppose an airline conducts a survey of six flights from Boston to Salt Lake City over Thanksgiving weekend to determine the number of babies on the flights. It determines the amount of safety equipment needed based on the results of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.
b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

16. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences with detailed description.

17. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences with detailed description.

18. List some practical difficulties involved in getting accurate results from a telephone survey.

19. List some practical difficulties involved in getting accurate results from a mailed survey.

20. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. What type of sampling did she use?

21. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly, and then every eighth house in the neighborhood around the park was interviewed. What was the sampling method?

22. Name the sampling method used in each of the following situations:

a. A woman in the airport hands out questionnaires to travelers, asking them to evaluate the airport's service. She does not stop travelers who are hurrying through the airport with their hands full of luggage, instead only asking travelers who are sitting near gates and not taking naps while they wait.
b. A teacher wants to know if her students are doing homework, so she randomly selects two rows and then calls on all students in those rows (row two and row five) to present the solutions to homework problems to the class.

c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires asking for information about their age and other variables of interest.
d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
e. A political party wants to know how voters are reacting to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked for whom they intend to vote and whether the debate changed their opinion of the candidates.

---

23. A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that, if they had $2,000 to spend, they would use it for computer equipment. In addition, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your "gut feeling," do you believe the percents accurately reflect the US population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?
c. Additional information: The survey, reported by the Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show, "America's Smithsonian." With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

---

24. The Well-Being Index is a survey that follows trends of US residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below. Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.[4]

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During how many of the last 30 days did poor health keep you from doing your usual activities?
c. On how many of the last seven days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

25. In advance of the 1936 presidential election, a magazine titled *Literary Digest* released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.[5]

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
b. What effect does the low response rate have on the reliability of the sample?
c. Are these problems examples of sampling error or non-sampling error?
d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. These researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population.[6] Quota sampling is an example of which sampling method described in this module?

---

26. Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's Uniform Crime Report. One analysis of this data found a strong connection between education and crime, indicating that higher levels of education in a community correspond to higher crime rates.[7]

Which of the potential problems with samples could explain this connection?

- Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.
- Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

---

27. YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

"Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?"[8]

As of April 25, 11 people had responded, answering "NO!"

Which of the potential problems with samples discussed in this module could explain this connection?

- Self-selected samples: Only people who are interested in the topic are choosing to respond.
- Sample size issues: A sample with only 11 participants will not accurately represent the opinions of a nation.
- Undue influence: The question is worded in a specific way to generate a specific response.
- Self-funded or self-interest studies: This question is generated to support one person's claim and is designed to get the answer that the person desires.

---

28. A scholarly article about response rates begins with the following quote:

"Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."[9]

The Pew Research Center for People and the Press admits:

"The percentage of people we interview—out of all we try to interview—has been declining over the past decade or more."[10]

a. What are some reasons for the decline in response rate over the past decade?
b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

---

29. During the 2010-2011 academic year, 771 distance learning students at Long Beach City College responded to surveys. Highlights of the summary report are listed in Figure 1.19.

Summary report	
Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

Figure 1.19

a. What percent of the students surveyed do not have a computer at home?
b. About how many students in the survey live at least 16 miles from campus?
c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

30. Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all or might get a delayed delivery if the book is backordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook from each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study. Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

References

Text

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).

"Vitamin E and Health," Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritionsource/vitamin-e (accessed May 1, 2013).

Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).

Ankita Mehta. "Daily Dose of Aspiring Helps Reduce Heart Attacks: Study," International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html (accessed May 1, 2013).

M.L. Jacskon et al., "Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors," Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).

"Earthquake Information by Year," U.S. Geological Survey. http://earthquake.usgs.gov/earthquakes/eqarchives/year (accessed May 1, 2013).

"Fatality Analysis Report Systems (FARS) Encyclopedia," National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).

Meier, Paul. “The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine.” *Statistics: a guide to the unknown*. San Francisco: Holden-Day (1972): 2-13.

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

“America’s Best Small Companies,” http://www.forbes.com/best-small-companies/list (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

“April 2013 Air Travel Consumer Report,” U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report (accessed May 1, 2013).

Lori Alden, “Statistics can be Misleading,” econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).

Maria de los A. Medina, “Ethics in Statistics,” Based on “Building an Ethics Module for Business, Science, and Engineering Students” by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest (accessed May 1, 2013).

McClung, M. Collins, D. “Because I know it will!”: placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology*. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

Yudhijit Bhattacharjee, “The Mind of a Con Man,” Magazine, New York Times, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).

“Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel,” Tillburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd-bce5-4385-b9ff-05b840caeae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).

Andrew Gelman, “Open Data and Open Methods,” Ethics and Statistics, http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President

“The Literary Digest Poll,” Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).

“Gallup Presidential Election Trial-Heat Trends, 1936–2008,” Gallup Politics http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus (accessed May 1, 2013).

Data from San Jose Mercury News

Notes

1. "Crash Test Dummies," The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).
2. Ankita Mehta, "Daily Dose of Aspirin Helps Reduce Heart Attacks: Study," *International Business Times*, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).
3. "U.S. Geological Survey, "Earthquake Information by Year," http://earthquake.usgs.gov/earthquakes/eqarchives/year (accessed May 1, 2013)
4. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx (accessed May 1, 2013).
5. Dominic Lusinchi, "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" *Social Science History* 36(1): 23-54, 2012, Also available online at https://www.jstor.org/stable/41407095 (accessed January 26, 2021).
6. Data from "How George Gallup Picked the President" on Book of Odds. Available online at http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President.
7. "United States: Uniform Crime Report – State Statistics from 1960–2011." The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).
8. lastbaldeagle. "On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes," 2013. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
9. Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," *Public Opinion Quarterly*, 70(5), 2006, http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
10. Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls (accessed May 1, 2013).

Chapter 2 Extra Practice

2.1 Descriptive Statistics and Frequency Distributions

1. What are the two types of descriptive statistical methods?

2.2 Displaying and Describing Categorical Distributions

1. What are the two basic options for graphing categorical data?

---

2. When describing categorical data we want to note what two aspects?

---

3. When describing the level of variability in categorical data, we want to think about it as ________.

2.3 Displaying Quantitative Distributions

1. Create a histogram for the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data since books are counted.

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
2, 2, 2, 2, 2, 2, 2, 2, 2, 2
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3
4, 4, 4, 4, 4, 4
5, 5, 5, 5, 5
6, 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value, and add 0.5 to 6, the largest data value. Then the starting point is 0.5, and the ending value is 6.5.

Next calculate the width of each bar or class interval. If the data are discrete, and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient.

If you are using an offline version of this text, access the activity using the QR code.

An interactive H5P element has been excluded from this version of the text. You can view it online here: *https://pressbooks.lib.vt.edu/significantstatistics/?p=1073#h5p-73*

Calculate the number of bars as follows:

$$\frac{6.5-0.5}{\text{numberofbars}} = 1$$

1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.

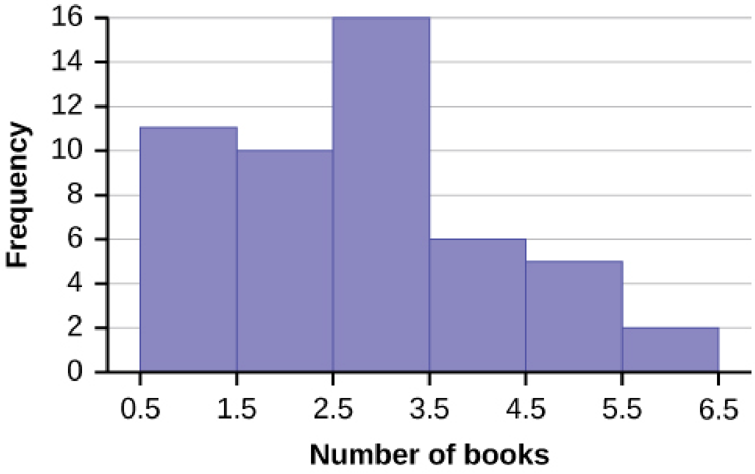


Figure 2.59. Figure description available at the end of the section.

2. We will construct an overlay frequency polygon comparing the scores from the figure below with the students' final numeric grades.

Lower bound	Upper bound	Frequency	Cumulative frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Figure 2.60: Frequency distribution for calculus final test scores

Lower bound	Upper bound	Frequency	Cumulative frequency
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100

Figure 2.61: Frequency distribution for calculus final test scores

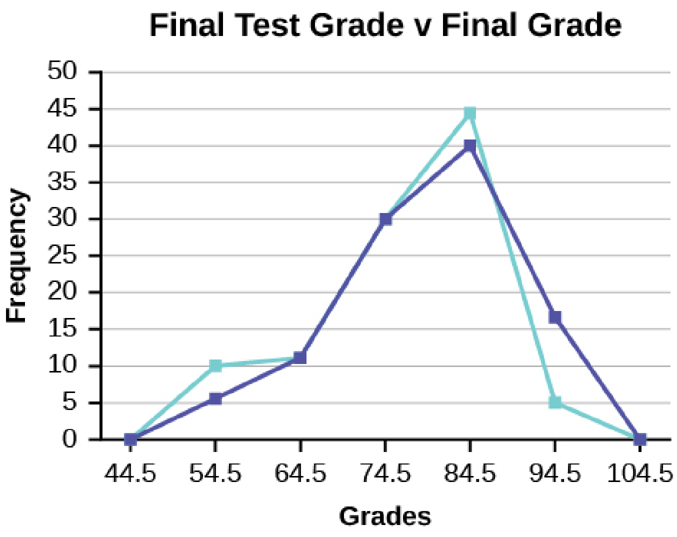


Figure 2.62. Figure description available at the end of the section.

3. Construct a frequency polygon of US Presidents' ages at inauguration shown in the figure below.[1]

Age at inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	3

Figure 2.63

4. Construct frequency polygons for the following datasets:

Pulse rates for women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

Figure 2.64

Actual speed in a 30 MPH zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

Figure 2.65

Tar (mg) in non-filtered cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Figure 2.66

---

5. Construct a frequency polygon from the frequency distribution for the 50 highest-ranked countries for depth of hunger.[2]

Depth of hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

Figure 2.67

---

6. Use the two frequency tables to compare the life expectancies of men and women from 20 randomly selected countries. Include an overlaid frequency polygon, and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?[3]

Life expectancy at birth (women)	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Figure 2.68

Life expectancy at birth (men)	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

Figure 2.69

7. The following table is a portion of a dataset from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.[4]

	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Figure 2.70

8. Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.[5]

	Female	Male	Total
1855	45,545	47,804	93,349
1856	49,582	52,239	101,821
1857	50,257	53,158	103,415
1858	50,324	53,694	104,018
1859	51,915	54,628	106,543
1860	51,220	54,409	105,629
1861	52,403	54,606	107,009
1862	51,812	55,257	107,069
1863	53,115	56,226	109,341

	Female	Male	Total
1864	54,959	57,374	112,333
1865	54,850	58,220	113,070
1866	55,307	58,360	113,667
1867	55,527	58,517	114,044
1868	56,292	59,222	115,514
1869	55,033	58,321	113,354
1870	56,431	58,959	115,390
1871	56,099	60,029	116,128
1872	57,472	61,293	118,765
1873	58,233	61,467	119,700
1874	60,109	63,602	123,711
1875	60,146	63,432	123,578

Figure 2.71

---

9. The following datasets list the number of full-time police officers per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan, during the period from 1961 to 1973.[6]

	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36	28.03	31.49	37.39	46.26	47.24	52.33

Figure 2.72

a. Construct a double time series graph using a common x-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

2.5 Measures of Location and Outliers

1. Test scores for a college statistics class held during the day are 99, 56, 78, 55.5, 32, 90, 80, 81, 56, 59, 45, 77, 84.5, 84, 70, 72, 68, 32, 79, and 90. Test scores for a college statistics class held during the evening are 98, 78, 68, 83, 81, 89, 88, 76, 65, 45, 98, 90, 80, 84.5, 85, 79, 78, 98, 90, 79, 81, and 25.5.[7]

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each dataset, what percentage of the data is between the smallest value and the first quartile? The first quartile and the median? The median and the third quartile? The third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

---

2. The following dataset shows the heights in inches for the boys in a class of 40 students: 66, 66, 67, 67, 68, 68, 68, 68, 68, 69, 69, 69, 70, 71, 72, 72, 72, 73, 73, 74. The following dataset shows the heights in inches for the girls in a class of 40 students: 61, 61, 62, 62, 63, 63, 63, 65, 65, 65, 66, 66, 66, 67, 68, 68, 68, 69, 69, 69. Construct a box plot using a graphing calculator for each dataset, and state which box plot has the wider spread for the middle 50% of the data.

---

3. Graph a box-and-whisker plot for the data values shown.

10, 10, 10, 15, 35, 75, 90, 95, 100, 175, 420, 490, 515, 515, 790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

---

4. Graph a box-and-whisker plot for the data values shown.

0, 5, 5, 15, 30, 30, 45, 50, 50, 60, 75, 110, 140, 240, 330

---

5. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars, 19 generally sell four cars, 12 generally sell five cars, 9 generally sell six cars, and 11 generally sell seven cars.

a. Construct a box plot. Use a ruler to measure and scale accurately.
b. Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or

concentrated in some areas but not in others? How can you tell?

6. In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetimes. The following box plots display the results.

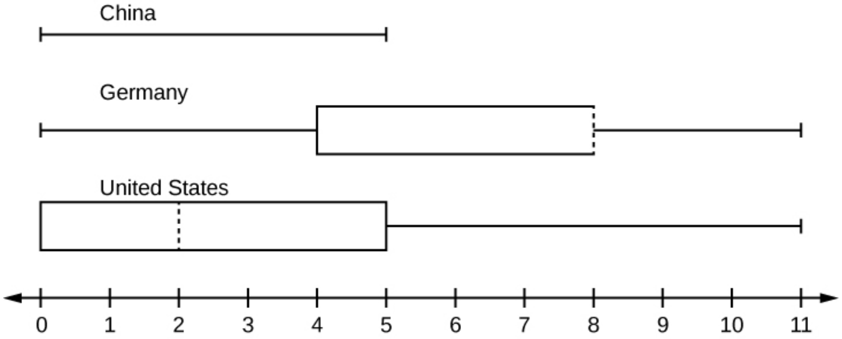


Figure 2.73. *Figure description available at the end of the section.*

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Have more Americans or more Germans surveyed been to over eight foreign countries?
c. Compare the three box plots. What do the comparisons imply about the foreign travel of 20-year-old residents of the three countries?

7. Given the following box plot, answer the questions.

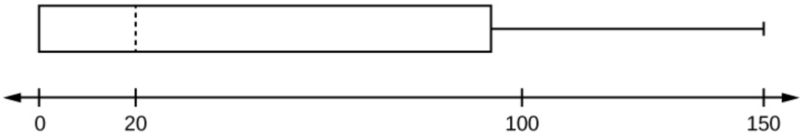


Figure 2.74. *Figure description available at the end of the section.*

a. Think of an example (in words) where the data might fit into the above box plot. In two-to-five sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second and third quartiles are far apart?

8. Given the following box plots, answer the questions.

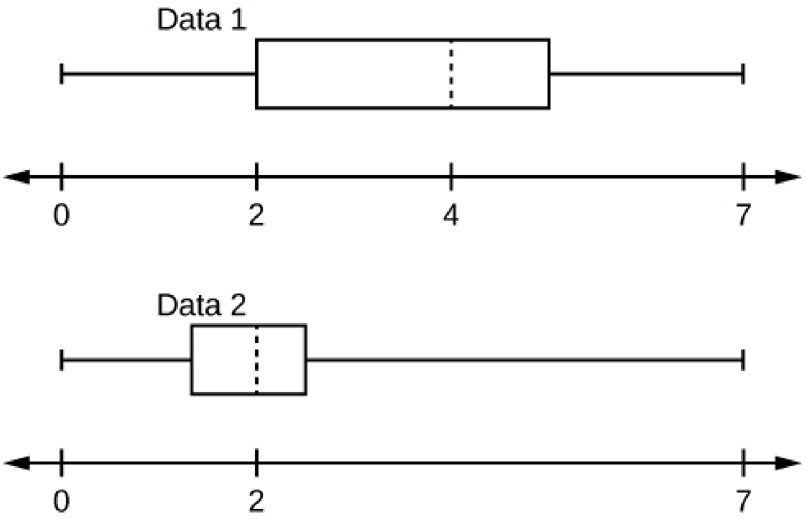


Figure 2.75. *Figure description available at the end of the section.*

a. In complete sentences, explain why each statement is false.
    i. Data 1 has more data values above two than Data 2 has above two.
    ii. The datasets cannot have the same mode.
    iii. For Data 1, there are more data values below four than there are above four.
b. For which group, Data 1 or Data 2, is the value of 7 more likely to be an outlier? Explain why in complete sentences.

---

9. A survey was conducted of 130 purchasers of new BMW 3 Series cars, 130 purchasers of new BMW 5 Series cars, and 130 purchasers of new BMW 7 Series cars. In it, people were asked the age they were when they purchased their cars. The following box plots display the results.

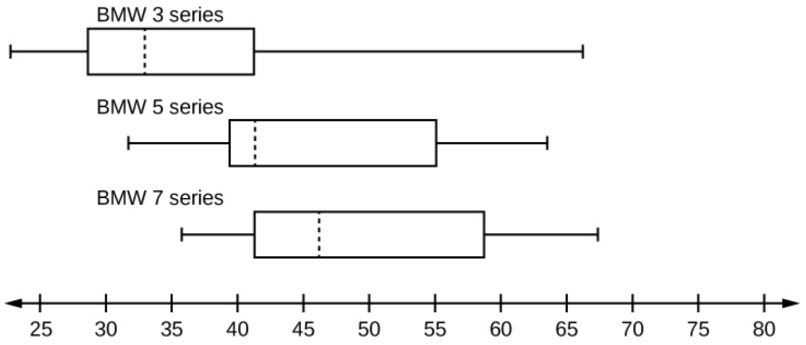


Figure 2.76. *Figure description available at the end of the section.*

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
b. Which group is most likely to have an outlier? Explain.
c. Compare the three box plots. What do they imply about ages when purchasing BMWs from different series?"
d. Look at the BMW 5 Series. Which quarter has the smallest spread of data? What is the spread?
e. Look at the BMW 5 Series. Which quarter has the largest spread of data? What is the spread?
f. Look at the BMW 5 Series. Estimate the interquartile range (IQR).
g. Look at the BMW 5 Series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
h. Look at the BMW 5 Series. Which interval has the fewest data in it? How do you know this?

    i. 31–35
    ii. 38–41
    iii. 41–64

---

10. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

Number of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Figure 2.77

Construct a box plot of the data.

---

11. Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Age group	Percent of community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1

Age group	Percent of community
55–64	11.9
65+	10.3

Figure 2.78

a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will *not* be the same width for this example. Why not? What impact does this have on the reliability of the graph?
b. What percentage of the community is under the age of 35?
c. Which box plot most resembles the information above?

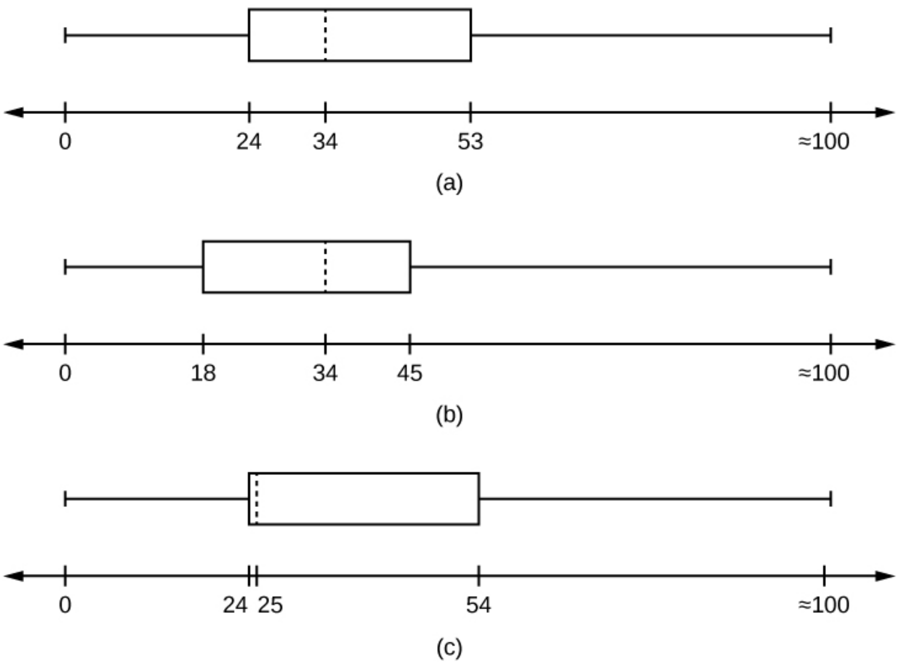


Figure 2.79. *Figure description available at the end of the section.*

12. For the following 13 real estate prices, calculate the IQR and determine if any prices are potential outliers. Prices are in dollars.

Data: 389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

13. For the following 11 salaries, calculate the IQR and determine if any salaries are outliers. The salaries are in dollars.

$33,000, $64,500, $28,000, $54,000, $72,000, $68,500, $69,000, $42,000, $54,000, $120,000, $40,500

---

14. For the two datasets about test scores in the first question of this section, find the following:

a. Both interquartile ranges. Compare the two.
b. Any outliers in either set.

---

15. Find the interquartile range for the following two datasets and compare them.

Test Scores for Class A:
69, 96, 81, 79, 65, 76, 83, 99, 89, 67, 90, 77, 85, 98, 66, 91, 77, 69, 80, 94

Test Scores for Class B:
90, 72, 80, 92, 90, 97, 92, 75, 79, 68, 70, 80, 99, 95, 78, 73, 71, 68, 95, 100

---

16. Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

Amount of sleep per school night (hours)	Frequency	Relative frequency	Cumulative relative frequency
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Figure 2.80

a. Find the 28th percentile.
b. Find the median.
c. Find the third quartile.

17. Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative frequency	Cumulative relative frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Figure 2.81

---

18. Using the table below:

Amount of sleep per school night (hours)	Frequency	Relative frequency	Cumulative relative frequency
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Figure 2.82

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

---

19. Refer to the table below. Find the third quartile. What is another name for the third quartile?

Amount of time spent on route (hours)	Frequency	Relative frequency	Cumulative relative frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Figure 2.83

20. Twenty-nine ages of winners of the Academy Award for Best Actor are listed below *in order from smallest to largest.*

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

a. Find the 40th percentile.
b. Find the 78th percentile.

---

21. Twenty-nine ages of winners of the Academy Award for Best Actor are listed below *in order from smallest to largest.*

18, 18, 21, 22, 25, 26, 27, 29, 30, 31, 31, 33, 36, 37, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

a. Find the percentile of 37.
b. Find the percentile of 72.

---

22. Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

---

23. For runners in a race, a *low time* means a faster run. The winners in a race have the shortest running times.

a. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
c. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

---

24. For runners in a race, a *higher speed* means a faster run.

a. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

25. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

26. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

27. In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

28. In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car sustained $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain, and write a one-sentence interpretation of the 90th percentile in the context of this problem.

29. The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called "eligible in the local context"), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are eligible in the local context?

30. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is in the 34th percentile. The 34th percentile of housing prices is $240,000 in the town to which you want to move. In this town, can you afford 34% of the houses or 66% of the houses?

31. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. 14 people answered that they generally sell three cars, 19 generally sell four cars, 12 generally sell five cars, 9 generally sell six cars, and 11 generally sell seven cars. Identify the following:

a. First quartile
b. Second quartile = median = 50th percentile
c. Third quartile
d. Interquartile range (IQR)
e. 10th percentile
f. 70th percentile

---

32. The median age for Black US citizens is 30.9 years; for White US citizens, it is 42.3 years.

a. Based upon this information, give two reasons why the Black median age could be lower than the White median age.
b. Does the lower median age for Black citizens necessarily mean that Black citizens die younger than White citizens? Why or why not?
c. How might it be possible for Black and White citizens to die at approximately the same age, even though the median age for White citizens is higher?

---

33. Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in the figure below.

Salary ($)	Relative frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

Figure 2.84

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data. Include left endpoint, but not the right endpoint.
    i. Should all bars have the same width, based on the data? Why or why not?
    ii. How should the < 20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40th and 80th percentiles.
e. Construct a bar graph of the data.

34. Given the following box plot:

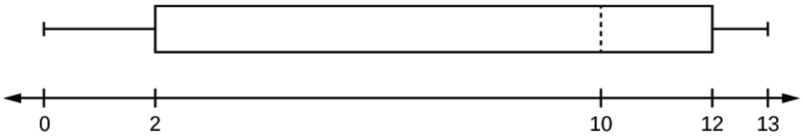


Figure 2.85. Figure description available at the end of the section.

a. Which quarter has the smallest spread of data? What is that spread?
b. Which quarter has the largest spread of data? What is that spread?
c. Find the interquartile range (IQR).
d. Are there more data in the 5-10 interval or in the 10-13 interval? How do you know this?
e. Which interval has the fewest data in it? How do you know this?

   i. 0–2
   ii. 2–4
   iii. 10–12
   iv. 12–13
   v. need more information

---

35. The following box plot shows the US population for 1990.

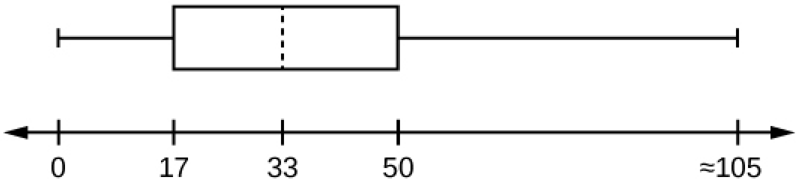


Figure 2.86. Figure description available at the end of the section.

a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
b. Of the population, 12.6% are age 65 and over. Approximately what percentage of the population are working age adults between the ages of 17 and 65?

---

36. On a 20-question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

37. On a 60-point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

---

38. At a community college, it was found that the 30th percentile of number of credit units for which students are enrolled is seven units. Interpret the 30th percentile in the context of this situation.

---

39. During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

---

40. Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. Note that a loss is shown by a negative weight gain.

Weight gain	Frequency
-2	3
-1	5
0	2
1	4
4	13
6	2
11	1

Figure 2.87

a. Calculate the following values:
   i. the average weight gain for the two weeks
   ii. the standard deviation
   iii. the first, second, and third quartiles
b. Construct a histogram and box plot of the data.

---

41. The figure below shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative frequency	Cumulative relative frequency
2.95–4.97	6	$\frac{6}{50} = 0.12$	0.12
4.97–6.99	7	$\frac{7}{50} = 0.14$	0.12 + 0.14 = 0.26

Rainfall (Inches)	Frequency	Relative frequency	Cumulative relative frequency
6.99–9.01	15	$\frac{15}{50} = 0.30$	0.26 + 0.30 = 0.56
9.01–11.03	8	$\frac{8}{50} = 0.16$	0.56 + 0.16 = 0.72
11.03–13.05	9	$\frac{9}{50} = 0.18$	0.72 + 0.18 = 0.90
13.05–15.07	5	$\frac{5}{50} = 0.10$	0.90 + 0.10 = 1.00
	Total = 50	Total = 1.00	

Figure 2.88

a. From the figure above, find the percentage of rainfall that is less than 9.01 inches.

b. Find the percentage of rainfall that is between 6.99 and 13.05 inches.

c. Find the number of towns that have rainfall between 2.95 and 9.01 inches.

d. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

---

42. Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2, 5, 7, 3, 2, 10, 18, 15, 20, 7, 10, 18, 5, 12, 13, 12, 4, 5, 10. The following table was produced:

Data	Frequency	Relative frequency	Cumulative relative frequency
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{3}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

Figure 2.89

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Fewer than 12 miles? Between five and 13 miles (not including five and 13 miles)?

---

43. Sixty adults with gum disease were asked the number of times per week they used to floss before being diagnosed. The (incomplete) results are shown in the figure below:

Times flossing per week	Frequency	Relative frequency	Cumulative relative frequency
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Figure 2.90

a. Fill in the blanks in the figure above.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

---

44. Nineteen immigrants to the US were asked how many years, to the nearest year, they have lived in the US. The data are as follows: 2, 5, 7, 2, 2, 10, 20, 15, 0, 7, 0, 20, 5, 12, 15, 12, 4, 5, 10.

Data	Frequency	Relative frequency	Cumulative relative frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895

Data	Frequency	Relative frequency	Cumulative relative frequency
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Figure 2.91

a. Fix the errors in the figure above. In addition, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the US for 5 years.”
c. Fix the statement in *b.* to make it correct.
d. What fraction of the people surveyed have lived in the US five or seven years?
e. What fraction of the people surveyed have lived in the US at most 12 years?
f. What fraction of the people surveyed have lived in the US fewer than 12 years?
g. What fraction of the people surveyed have lived in the US from five to 20 years, inclusive?

---

45. The population in Park City is made up of children, working-age adults, and retirees. The figure below shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Figure 2.92

---

46. The following data are the distances (in kilometers) from a home to local supermarkets:

1.1, 1.5, 2.3, 2.5, 2.7, 3.2, 3.3, 3.3, 3.5, 3.8, 4.0, 4.2, 4.5, 4.5, 4.7, 4.8, 5.5, 5.6, 6.5, 6.7, 12.3

a. Create a stemplot using the data.
b. Do the data seem to have any concentration of values?

47. The following data show the distances (in miles) from the homes of off-campus statistics students to the college:

0.5, 0.7, 1.1, 1.2, 1.2, 1.3, 1.3, 1.5, 1.5, 1.7, 1.7, 1.8, 1.9, 2.0, 2.2, 2.5, 2.6, 2.8, 2.8, 2.8, 3.5, 3.8, 4.4, 4.8, 4.9, 5.2, 5.5, 5.7, 5.8, 8.0

Create a stem plot using the data, and identify any outliers.

---

48. For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32, 32, 33, 34, 38, 40, 42, 42, 43, 44, 46, 47, 47, 48, 48, 48, 49, 50, 50, 51, 52, 52, 52, 53, 54, 56, 57, 57, 60, 61

Construct a stem plot for the data.

---

49. The table below shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969-1970	39	43	1990–1991
46	36	1970-1971	44	38	1991–1992
46	36	1971-1972	39	43	1992–1993
36	46	1972-1973	25	57	1993–1994
47	35	1973-1974	40	42	1994–1995
51	31	1974-1975	36	46	1995–1996
53	29	1975-1976	26	56	1996–1997
51	31	1976-1977	32	50	1997–1998
41	41	1977-1978	19	31	1998–1999
36	46	1978-1979	54	28	1999–2000
32	50	1979-1980	57	25	2000–2001
51	31	1980-1981	49	33	2001–2002
40	42	1981-1982	47	35	2002–2003
39	43	1982-1983	54	28	2003–2004
42	40	1983-1984	69	13	2004–2005
48	34	1984-1985	56	26	2005–2006
32	50	1985-1986	52	30	2006–2007
25	57	1986-1987	45	37	2007–2008

Losses	Wins	Year	Losses	Wins	Year
32	50	1987-1988	35	47	2008–2009
30	52	1988-1989	29	53	2009–2010

Figure 2.93

---

50. In a survey, 40 people were asked how many times per year their car was in the shop for repairs. The results are shown in the table below. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Figure 2.94

---

51. Using the following dataset, construct a histogram.

Number of hours my classmates spent playing video games on weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Figure 2.95

---

52. The following data represent the number of employees at various restaurants in New York City:

22, 35, 15, 26, 40, 28, 18, 20, 25, 34, 39, 42, 24, 22, 19, 27, 22, 34, 40, 20, 38, 28

Using this data, create a histogram. Use 10–19 as the first interval.

53. Suppose 111 people who shopped in a special T-shirt store were asked the number of T-shirts they own that cost more than $19 each.

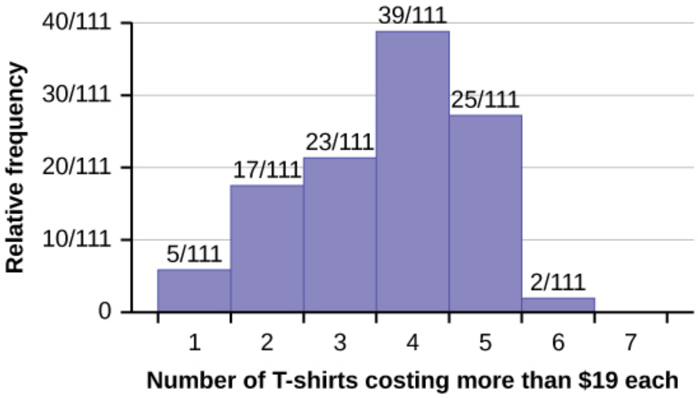


Figure 2.96. *Figure description available at the end of the section*.

a. The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21
b. 59
c. 41
d. Cannot be determined

b. If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

a. cluster
b. simple random
c. stratified
d. convenience

54. Following are the 2010 obesity rates of the US states and Washington, DC.[8]

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5

State	Percent (%)	State	Percent (%)	State	Percent (%)
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Figure 2.97

Construct a bar graph of obesity rates of your state and the four states closest to your state, labeling the x-axis with the states. Answers will vary.

---

55. Student grades on a chemistry exam were 77, 78, 76, 81, 86, 51, 79, 82, 84, and 99.

a. Construct a stem-and-leaf plot of the data.
b. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

---

56. The table below contains the 2010 obesity rates of US states and Washington, DC.[9]

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5

State	Percent (%)	State	Percent (%)	State	Percent (%)
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Figure 2.98

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."
c. Construct a bar graph for all the states beginning with the letter "M."

---

57. For each of the following datasets, create a stem plot and identify any outliers.

a. The miles per gallon rating for 30 cars are shown below (lowest to highest).
19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43
b. The height (in feet) of 25 trees is shown below (lowest to highest).
25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54
c. The data are the prices of different laptops at an electronics store. Round each value to the nearest ten.
249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610
d. The data are daily high temperatures in a town for one month.
61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

---

58. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. The figure below shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%

Seasons	Number of students	Proportion of population
Winter	6	18%

Figure 2.99

Using the data from Ms. Ramirez's math class, construct a bar graph showing the percentages.

59. David County has six high schools. Each school sent students to participate in a county-wide science competition. The figure below shows the percentage breakdown of competitors from each school and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High school	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Figure 2.100

Use the data from the David County science competition supplied above to construct a bar graph that shows the county-wide population percentage of students at each school.

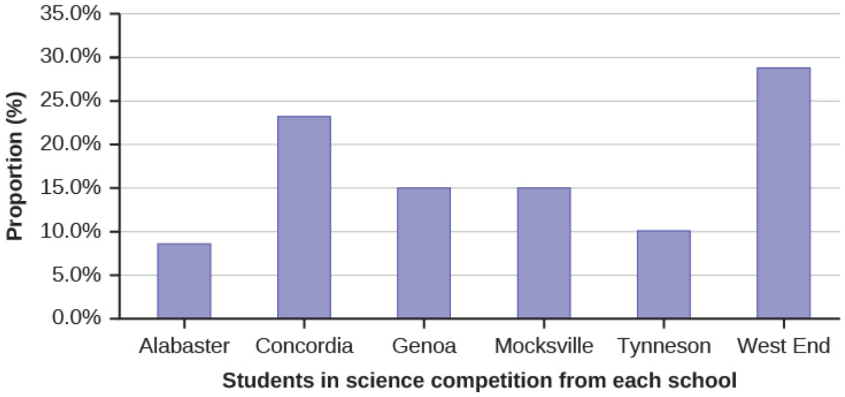


Figure 2.101. Figure description available at the end of the section.

2.6 Measures of Center

1. The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3, 4, 5, 7, 7, 7, 7, 8, 8, 9, 9, 10, 10, 10, 10, 10, 11, 12, 12, 13, 14, 14, 15, 15, 17, 17, 18, 19, 19, 19, 21, 21, 22, 22, 23, 24, 24, 24, 24

---

2. In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the "center," the mean or the median?

---

3. The number of books checked out from the library from 25 students are as follows:

0, 0, 0, 1, 2, 3, 3, 4, 4, 5, 5, 7, 7, 7, 7, 8, 8, 8, 9, 10, 10, 11, 11, 12, 12

Find the mode.

---

4. Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

Figure 2.102

b.

Daily low temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

Figure 2.103

c.

Points per game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Figure 2.104

---

5. The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16, 17, 19, 20, 20, 21, 23, 24, 25, 25, 25, 26, 26, 27, 27, 27, 28, 29, 30, 32, 33, 33, 34, 35, 37, 39, 40

a. Calculate the mean.
b. Identify the median.
c. Identify the mode.

---

6. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. 14 people answered that they generally sell three cars, 19 generally sell four cars, 12 generally sell five cars, 9 generally sell six cars, and 11 generally sell seven cars. Calculate the following:

a. Calculate the sample mean.
b. Identify the median.
c. Identify the mode.

---

7. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.[10]

Percent of population obese	Number of countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1

Percent of population obese	Number of countries
65.45–74.45	0
74.45–83.45	1

Figure 2.105

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

---

8. The following figure gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?[11]

Percent of children underweight	Number of countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Figure 2.106

The mean percentage, $\overline{x} = \frac{1328.65}{50} = 26.75$

---

9. Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

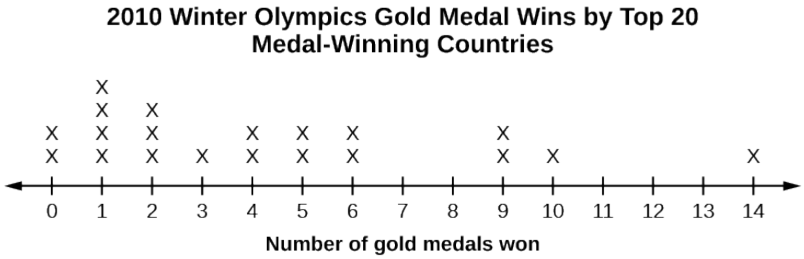


Figure 2.107. Figure description available at the end of the section.

b.

The ages former US presidents died	
4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

Figure 2.108

c.

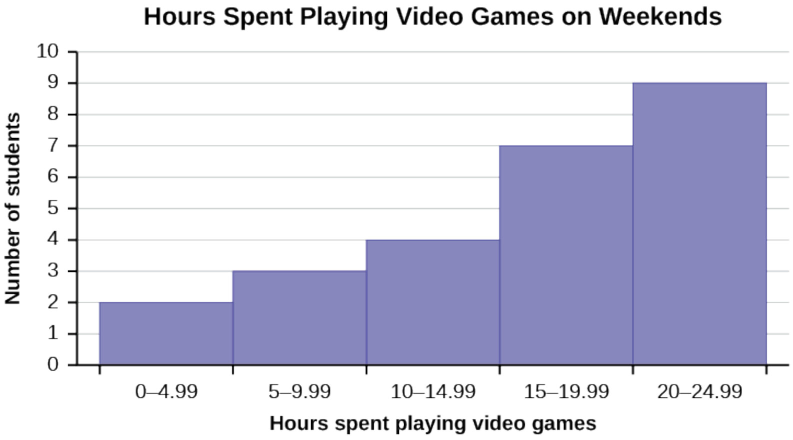


Figure 2.109. Figure description available at the end of the section.

10. State whether the data are symmetrical, skewed to the left, or skewed to the right.

a. 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5
b. 16, 17, 19, 22, 22, 22, 22, 22, 23
c. 87, 87, 87, 87, 87, 88, 89, 89, 90, 91

11. When the data are skewed left, what is the typical relationship between the mean and median?

12. When the data are symmetrical, what is the typical relationship between the mean and median?

13. What word describes a distribution that has two modes?

---

14. Use the following graph to answer a.-c.

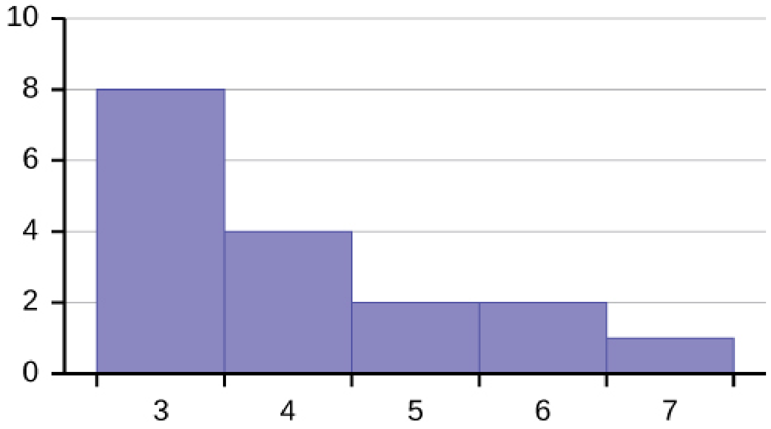


Figure 2.110. *Figure description available at the end of the section*.

a. Describe the shape of this distribution.
b. Describe the relationship between the mode and the median of this distribution.
c. Describe the relationship between the mean and the median of this distribution.

---

15. Data: 11, 11, 12, 12, 12, 12, 13, 15, 17, 22, 22, 22

a. Is the data perfectly symmetrical? Why or why not?
b. Which is the largest, the mean, the mode, or the median of the dataset?

---

16. Data: 56, 56, 56, 58, 59, 60, 62, 64, 64, 65, 67

a. Is the data perfectly symmetrical? Why or why not?
b. Which is the largest, the mean, the mode, or the median of the dataset?

---

17. Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

18. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

---

19. The median age of the US population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. Does the median age rising mean the actual age of children in 1991 was less than in 1980? Why or why not?

---

20. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\overline{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

Figure 2.111

a. How can you determine which survey was correct?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the following two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

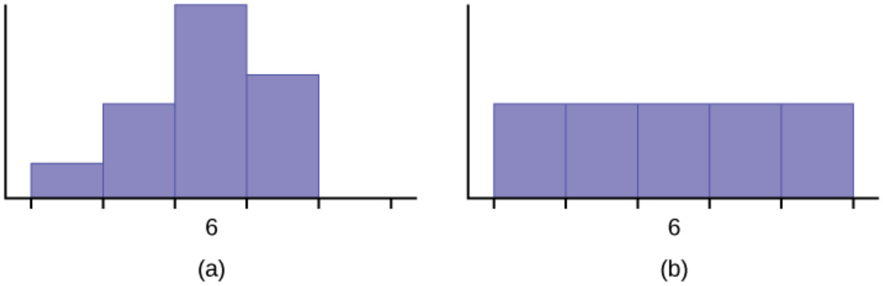


Figure 2.112. Figure description available at the end of the section.

d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

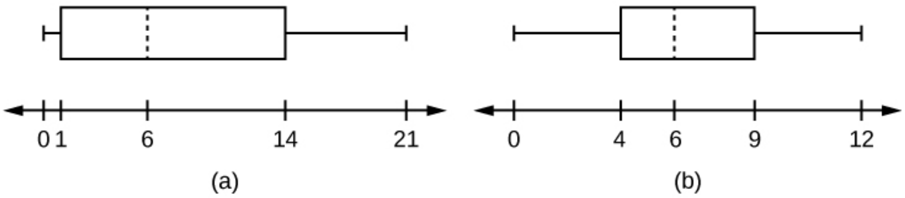


Figure 2.113. *Figure description available at the end of the section*.

---

21. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency
7	1
14	3
15	1
18	1
19	4
20	3
22	1
23	1
26	1
40	2
42	2
	Total = 20

Figure 2.114

What is the IQR?

a. 11
b. 35
c. 15
d. 8

What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Is this a sample or the entire population?

a. sample
b. entire population
c. neither

---

22. How much time does it take to travel to work? The figure below shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

Mean commute times (by state)									
24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

Figure 2.115

---

23. Find the midpoint for each class. These will be graphed on the x-axis. The frequency values will be graphed on the y-axis values.

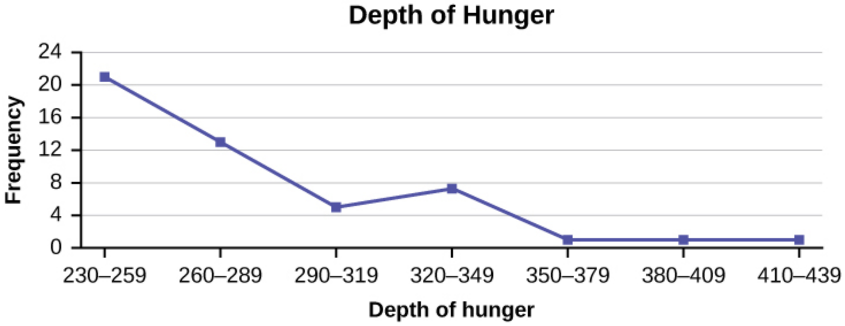


Figure 2.116. Figure description available at the end of the section.

2.7 Measures of Spread

1. Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place:
   i. The sample mean
   ii. The sample standard deviation
   iii. The median
   iv. The first quartile
   v. The third quartile
   vi. IQR
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

---

2. The following data show the different types of pet food that stores in the area carry:

6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12

Calculate the sample mean and the sample standard deviation to one decimal place.

---

3. The following data are the distances (in miles) between 20 retail stores and a large distribution center:

29, 37, 38, 40, 58, 67, 68, 69, 76, 86, 87, 95, 96, 96, 99, 106, 112, 127, 145, 150

a. Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.
b. Find the value that is one standard deviation below the mean.

---

4. Fredo and Karl, two baseball players on different teams, wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball player	Batting average	Team batting average	Team standard deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Figure 2.117

For Fredo:

$$z = \frac{0.158 - 0.166}{0.012} = -0.67$$

For Karl:

$$z = \frac{0.177 - 0.189}{0.015} = -0.8$$

Fredo's z-score of −0.67 is higher than Karl's z-score of −0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Use the table above to find the value that is three standard deviations (a) above the mean and (b) below the mean.

---

5. Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

Figure 2.118

Daily low temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

Figure 2.119

Points per game	Frequency
49.5–59.5	14

Points per game	Frequency
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Figure 2.120

---

6. The population parameters below describe the number of full-time equivalent students (FTES) each year at ABC University from 1976–1977 through 2004–2005.

μ = 1,000 FTES

Median = 1,014 FTES

σ = 474 FTES

First quartile = 528.5 FTES

Third quartile = 1,447.5 FTES

n = 29 years

a. A sample of 11 years is taken. About how many are expected to have a FTES of 1,014 or above? Explain how you determined your answer.
b. 75% of all years have an FTES:
   i. At or below: ______
   ii. At or above: ______
c. What is the population standard deviation?
d. What percent of the FTES were from 528.5 to 1447.5? How do you know?
e. What is the IQR? What does the IQR represent?
f. How many standard deviations away from the mean is the median? *Additional Information:* The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here:
   FTES population **Year** 2005-06 2006–07 2007–08 2008–09 2009–10 2010–11 **Total FTFS** 1,585 1,690 1,735 1,935 2,021 1,890 *Figure 2.121*
g. Calculate the mean, median, standard deviation, the first quartile, the third quartile and the IQR. Round to one decimal place.
h. What additional information is needed to construct a box plot for the FTES for 2005-2006 through 2010-2011 and a box plot for the FTES for 1976-1977 through 2004-2005?
i. Compare the IQR for the FTES for 1976–77 through 2004–2005 with the IQR for the FTES for 2005-2006 through 2010–2011. Why do you suppose the IQRs are so different? Hint: Think about the number of

years covered by each time period and what happened to higher education during those periods.

---

7. Three students are applying to the same graduate school. They come from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School average GPA	School standard deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Figure 2.122

---

8. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type? Justify your answer.

---

9. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran one mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to their class? Explain why.

---

10. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in figure 2.123.[12]

Percent of population obese	Number of countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

Figure 2.123

a. What is the best estimate of the average obesity percentage for these countries?
b. What is the standard deviation for the listed obesity rates?
c. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
d. How “unusual” is the United States’ obesity rate compared to the average rate? Explain.

---

11. The figure below gives the percent of children under five considered to be underweight.[13]

Percent of children underweight	Number of countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Figure 2.124

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

---

12. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are shown in figure 2.125.

Number of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Figure 2.125

a. Find the sample mean, $\overline{x}$.
b. Find the approximate sample standard deviation, s.

---

13. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X represent the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

Figure 2.126

a. Find the sample mean, $\overline{x}$.
b. Find the sample standard deviation, s.
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40th percentile.
k. Find the 90th percentile.
l. Construct a line graph of the data.
m. Construct a stemplot of the data.

14. Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year:

177, 205, 210, 210, 232, 205, 185, 185, 178, 210, 206, 212, 184, 174, 185, 242, 188, 212, 215, 247, 241, 223, 220, 260, 245, 259, 278, 270, 280, 295, 275, 285, 290, 272, 273, 280, 285, 286, 200, 215, 185, 230, 250, 241, 190, 260, 250, 302, 265, 290, 276, 228, 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from ________ to _______.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:
    i. the population mean, μ.
    ii. the population standard deviation, σ.
    iii. the weight that is two standard deviations below the mean.
j. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
k. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

---

15. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3, 8, −1, 2, 0, 5, −3, 1, −1, 6, 5, −2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

16. Refer to the figures below and determine which of the following (a.-d.) are true and which are false. Explain your solution to each part in complete sentences.

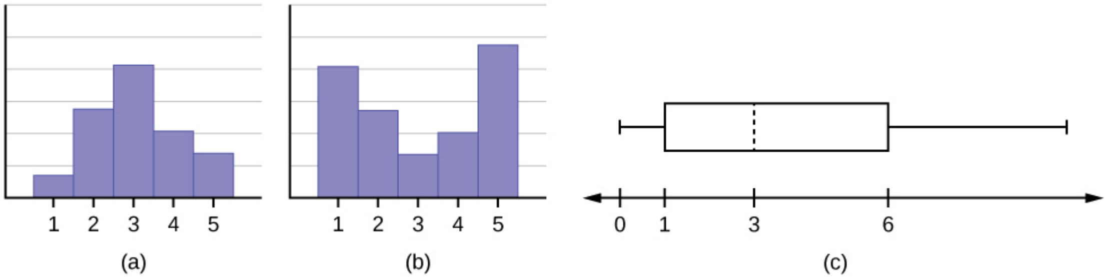


Figure 2.127. Figure description available at the end of the section.

a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for Graph (b) is larger than the standard deviation for Graph (a).
d. We cannot determine if any of the third quartiles for the three graphs is different.

---

17. In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X represent the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65th percentile.
d. Find the 10th percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from ________ days to _______ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference, mean, median, or mode? Explain.
k. Give two reasons why you think that three-to-five days seem to be popular lengths for engineering conferences.

18. A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6,414, 1,550, 2,109, 9,350, 21,828, 4,300, 5,944, 5,722, 2,825, 2,044, 5,481, 5,200, 5,853, 2,750, 10,012, 6,357, 27,000, 9,414, 7,681, 3,200, 17,500, 9,200, 7,380, 18,314, 6,557, 13,713, 17,768, 7,493, 2,771, 2,861, 1,263, 7,285, 28,165, 5,080, 11,622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."

b. Construct a histogram of the data.

c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?

d. Calculate the sample mean.

e. Calculate the sample standard deviation.

f. A school with an enrollment of 8,000 would be how many standard deviations away from the mean?

---

19. Let X represent the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Figure 2.128

a. What is the 80th percentile?

- 5
- 80
- 3
- 4

b. The number that is 1.5 standard deviations BELOW the mean is approximately _________.

- 0.7
- 4.8
- −2.8
- Cannot be determined

20. Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the figure below.

Number of books	Frequency	Relative frequency
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Figure 2.129

a. Are there any outliers in the data? Use an appropriate numerical test involving the IQR to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a. and c. of this problem give the same answer?
e. Examine the shape of the data. Which part, a. or c., of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data, mean, median or mode?

---

21. This figure contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total number of deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712

Year	Total number of deaths
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Figure 2.130

Answer each of the following questions and check your answers below.

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?
f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

---

22. The following figure contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total number of crashes	Year	Total number of crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	**Total**	**653,782**
2003	38,477		

Figure 2.131

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

---

23. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Number of courses	Frequency	Relative frequency	Cumulative relative frequency
1	30	0.6	
2	15		
3			

Figure 2.132

Fill in the blanks in the figure above.

a. What percent of students take exactly two courses?
b. What percent of students take one or two courses?

---

24. *Forbes* magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, had a stock price of at least $5 per share, and had reported annual revenue between $5 million and $1 billion. The figure below shows the ages of the chief executive officers for the top 60 ranked firms.

Age	Frequency	Relative frequency	Cumulative relative frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

Figure 2.133

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?

d. What is the cumulative relative frequency for CEOs younger than 55?

e. Which graph shows the relative frequency, and which shows the cumulative relative frequency?

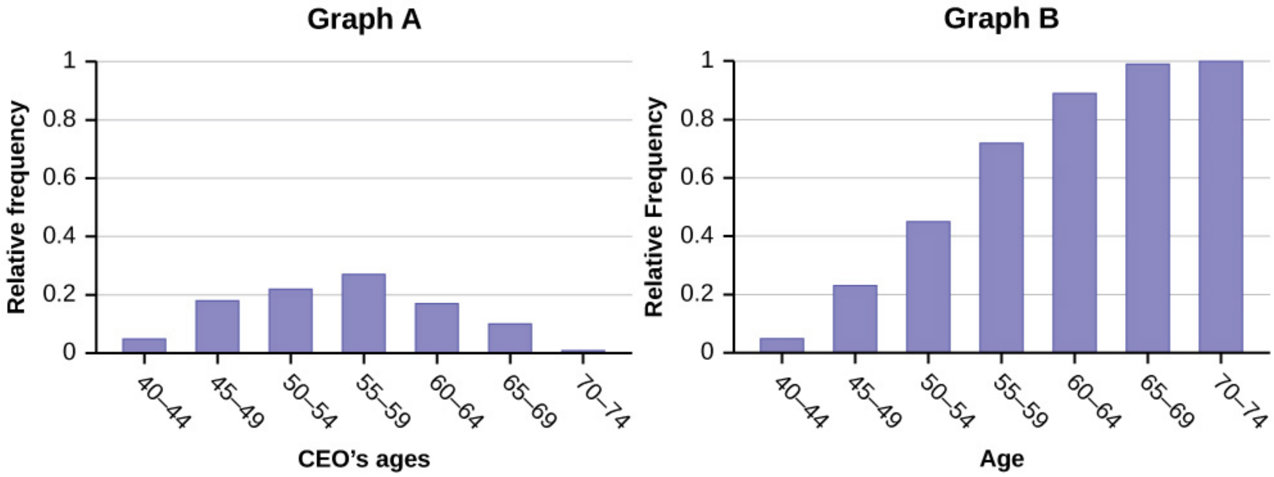


Figure 2.134. Figure description available at the end of the section.

25. The figure below contains data on hurricanes that have made direct hits on the US between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	Number of direct hits	Relative frequency	Cumulative frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Figure 2.135

a. What is the relative frequency of direct hits that were Category 4 hurricanes?

- 0.0768
- 0.0659
- 0.2601
- Not enough information to calculate

b. What is the relative frequency of direct hits that were AT MOST a Category 3 storm?

- 0.3480
- 0.9231
- 0.2601
- 0.3370

26. The following data are the shoe sizes of 50 male students. The sizes are discrete data since shoe size is measured in whole and half units only. Construct a histogram, and calculate the width of each bar or class interval, supposing you choose six bars.

9, 9, 9.5, 9.5, 10, 10, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5

11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5, 11.5, 11.5, 11.5, 11.5

12, 12, 12, 12, 12, 12, 12, 12.5, 12.5, 12.5, 12.5, 14

27. The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1

2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2

3, 3, 3, 3, 3, 3, 3, 3

20 student athletes play one sport. 22 student athletes play two sports. 8 student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

28. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. 14 people answered that they generally sell three cars, 19 generally sell four cars, 12 generally sell five cars, 9 generally sell six cars, and 11 generally sell seven cars. Complete the table.

Data value (number of cars)	Frequency	Relative frequency	Cumulative relative frequency

Figure 2.136

a. What does the frequency column sum to? Why?
b. What does the relative frequency column sum to? Why?
c. What is the difference between relative frequency and frequency for each data value?
d. What is the difference between cumulative relative frequency and relative frequency for each data value?
e. To construct the histogram for the data, determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

---

29. Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

Number of books	Frequency	Relative frequency
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Figure 2.137: Publisher A

Number of books	Frequency	Relative frequency
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Figure 2.138: Publisher B

Number of books	Frequency	Relative frequency
0-1	20	
2-3	35	
4-5	12	
6-7	2	
8-9	1	

Figure 2.139: Publisher C

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

30. Often, cruise ships conduct all on-board transactions (with the exception of gambling) on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. The following table is a summary of the bills for each group.

Amount ($)	Frequency	Relative frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Figure 2.140: Singles

Amount ($)	Frequency	Relative frequency
100–150	5	

Amount ($)	Frequency	Relative frequency
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

Figure 2.141: Couples

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the x-axis by $50 widths. Use relative frequency on the y-axis.
c. Construct a histogram for the couples group. Scale the x-axis by $50 widths. Use relative frequency on the y-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph by hand for the couples. Since each couple is paying for two individuals, instead of scaling the x-axis by $50, scale it by $100. Use relative frequency on the y-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that single individuals spend the same amount, more, or less than individuals who are part of a couple? Explain why in one or two complete sentences.

---

31. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are shown in Figure 2.142.

Number of movies	Frequency	Relative frequency	Cumulative relative frequency
0	5		

Number of movies	Frequency	Relative frequency	Cumulative relative frequency
1	9		
2	6		
3	4		
4	1		

Figure 2.142

a. Construct a histogram of the data.
b. Complete the columns of the chart.

---

32. Use the data to construct a line graph.

a. In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown below.

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

Figure 2.143

b. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown below.

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

Figure 2.144

c. Several children were asked how many TV shows they watch each day. The results of the survey are shown below.

Number of TV shows	Frequency
0	12

Number of TV shows	Frequency
1	18
2	36
3	7
4	2

Figure 2.145

Figure Descriptions

Figure 2.59: Histogram consists of six bars with the y-axis in increments of two from zero through 16 and the x-axis in intervals of one from 0.5-6.5. Bars taper off to the right.

Figure 2.62: This is an overlay frequency polygon that matches the supplied data. The x-axis shows the grades, and the y-axis shows the frequency. Both lines follow the same pattern, peaking around grade=84.5.

Figure 2.73: This shows three box plots graphed over a number line from zero to 11. The box plots match the supplied data, and compare the countries' results. The China box plot has a single whisker from zero to five. The Germany box plot's median is equal to the third quartile, so there is a dashed line at right edge of box. The America box plot does not have a left whisker.

Figure 2.74: This is a box plot graphed over a number line from zero to 150. There is no first, or left, whisker. The box starts at the first quartile, zero, and ends at the third quartile, 80. A vertical, dashed line marks the median, 20. The second whisker extends the third quartile to the largest value, 150.

Figure 2.75: This shows two box plots graphed over number lines from zero to seven. The first whisker in the data one box plot extends from zero to two. The box begins at the firs quartile, two, and ends at the third quartile, five. A vertical, dashed line marks the median at four. The second whisker extends from the third quartile to the largest value, seven. The first whisker in the data two box plot extends from zero to 1.3. The box begins at the first quartile, 1.3, and ends at the third quartile, 2.5. A vertical, dashed line marks the medial at two. The second whisker extends from the third quartile to the largest value, seven.

Figure 2.76: This shows three box plots graphed over a number line from 25 to 80. The first whisker on the BMW 3 plot extends from 25 to 30. The box begins at the firs quartile, 30 and ends at the third quartile, 41. A vertical, dashed line marks the median at 34. The second whisker extends from the third quartile to 66. The first whisker on the BMW 5 plot extends from 31 to 40. The box begins at the firs quartile, 40, and ends at the third quartile, 55. A vertical, dashed line marks the median at 41. The second whisker extends from 55 to 64. The first whisker on the BMW 7 plot extends from 35 to 41. The box begins at the first quartile, 41, and ends at the third quartile, 59. A vertical, dashed line marks the median at 46. The second whisker extends from 59 to 68.

Figure 2.79: Three box plots with values between 0 and 100. Plot one has Q1 at 24, M at 34, and Q3 at 53; Plot two has Q1 at 18, M at 34, and Q3 at 45; Plot three has Q1 at 24, M at 25, and Q3 at 54.

Figure 2.85: This is a horizontal box plot graphed over a number line from zero to 13. The first whisker

extends from the smallest value, zero, to the first quartile, two. The box begins at the first quartile and extends to third quartile, 12. A vertical, dashed line is drawn at median, 10. The second whisker extends from the third quartile to largest value, 13.

Figure 2.86: A box plot with values from zero to 105, with Q1 at 17, M at 33, and Q3 at 50.

Figure 2.96: A histogram showing the results of a survey. Of 111 respondents, five own one t-shirt costing more than $19, 17 own two, 23 own three, 39 own four, 25 own five, two own six, and no respondents own seven.

Figure 2.101: This is a bar graph that matches the supplied data. The x-axis shows the county high schools, and the y-axis shows the proportion of county students. Alabaster: 9%, Concordia: 24%, Genoa: 15%, Mocksville: 15%, Tynneson: 10%, West End: 29%

Figure 2.107: This dot plot matches the supplied data. The plot uses a number line from zero to 14. It shows two x's over zero, four x's over one, three x's over two, one x over three, two x's over the numbers four, five, six, and nine, and one x each over 10 and 14. There are no x's over the numbers seven, eight, 11, 12, and 13.

Figure 2.109: This is a histogram titled Hours Spent Playing Video Games on Weekends. The x-axis shows the number of hours spent playing video games with bars showing values at intervals of five. The y-axis shows the number of students. The first bar for 0 – 4.99 hours has a height of two. The second bar from 5 – 9.99 has a height of three. The third bar from 10 – 14.99 has a height of four. The fourth bar from 15 – 19.99 has a height of seven. The fifth bar from 20 – 24.99 has a height of nine.

Figure 2.110: This is a histogram which consists of five adjacent bars with the x-axis split into intervals of one from three to seven. The bar heights peak at the first bar and taper lower to the right.

Figure 2.112: This shows two histograms. The first histogram shows a fairly symmetrical distribution with a mode of 6. The second histogram shows a uniform distribution.

Figure 2.113: This shows two horizontal box plots. The first box plot is graphed over a number line from zero to 21. The first whisker extends from zero to one. The box begins at the first quartile, one, and ends at the third quartile, 14. A vertical, dashed line marks the median at six. The second whisker extends from the third quartile to the largest value, 21. The second box plot is graphed over a number line from zero to 12. The first whisker extends from zero to four. The box begins at the first quartile, four, and ends at the third quartile, nine. A vertical, dashed line marks the median at six. The second whisker extends from the third quartile to the largest value, 12.

Figure 2.116: This is a frequency polygon that matches the supplied data. The x-axis shows the depth of hunger, and the y-axis shows the frequency. Values decrease, slightly increase, and then decrease and flatten out.

Figure 2.127: This shows three graphs. The first is a histogram with a mode of three and fairly symmetrical distribution between one (minimum value) and five (maximum value). The second graph is a histogram with peaks at one (minimum value) and five (maximum value) with three having the lowest frequency. The third graph is a box plot. The first whisker extends from zero to one. The box begins at the firs quartile, one, and

ends at the third quartile, six. A vertical, dashed line marks the median at three. The second whisker extends from six on.

Figure 2.134: Graph A is a bar graph with seven bars. The x-axis shows CEO's ages in intervals of five years starting with 40 – 44. The y-axis shows the relative frequency in intervals of 0.2 from zero to one. The highest relative frequency shown is 0.27. Graph B is a bar graph with seven bars. The x-axis shows CEO's ages in intervals of five years starting with 40 – 44. The y-axis shows relative frequency in intervals of 0.2 from zero to one. The highest relative frequency shown is one.

References

Figures

Figure 2.59: Figure 2.6 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/2-2-histograms-frequency-polygons-and-time-series-graphs

Figure 2.62: Figure 2.9 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-2-histograms-frequency-polygons-and-time-series-graphs

Figure 2.73: Figure 2.45 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-homework

Figure 2.74: Figure 2.46 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-homework

Figure 2.75: Figure 2.47 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-homework

Figure 2.76: Figure 2.46 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/2-homework

Figure 2.79: Figure 2.47 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/2-bringing-it-together-homework

Figure 2.85: Figure 2.43 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-homework

Figure 2.86: Figure 2.44 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-homework

Figure 2.96: Figure from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-business-statistics/pages/2-homework

Figure 2.101: Figure 2.54 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/2-solutions

Figure 2.107: Figure 2.24 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode

Figure 2.109: Figure 2.25 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode

Figure 2.110: Figure 2.7.9 from LibreTexts Introductory Statistics (2020) (CC BY 4.0). Retrieved from https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(OpenStax)/02%3A_Descriptive_Statistics/2.07%3A_Skewness_and_the_Mean_Median_and_Mode

Figure 2.112: Figure 2.9.1 from LibreTexts Introductory Business Statistics (2020) (CC BY 4.0). Retrieved from https://biz.libretexts.org/Courses/Gettysburg_College/MGT_235%3A_Introductory_Business_Statistics/02%3A_Descriptive_Statistics/2.09%3A_Homework

Figure 2.113: Figure 2.51 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-bringing-it-together-homework

Figure 2.116: Figure 2.58 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-solutions

Figure 2.127: Figure 2.52 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/2-bringing-it-together-homework

Figure 2.134: Figure 1.11 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-business-statistics/pages/1-homework

Text

“State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau. http://quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).

“Table 5: Direct hits by mainland United States Hurricanes (1851-2004),” National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).

“Levels of Measurement,” http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

David Lane. “Levels of Measurement,” Connexions, http://cnx.org/content/m10809/latest (accessed May 1, 2013).

Dekker, Marcel. Data on annual homicides in Detroit, 1961–73 in Gunst & Mason, *Regression Analysis and its Application.*

“Timeline: Guide to the U.S. Presidents: Information on every president’s birthplace, political party, term of office, and more.” Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).

"Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

"Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at https://web.archive.org/web/20130917120224/https://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at http://www.ken-burbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/ (accessed August 21, 2013).

"9th Annual AP Report to the Nation." CollegeBoard, 2013. Available online at http://apreport.college-board.org/goals-and-findings/promoting-equity (accessed September 13, 2013).

Data from *West Magazine*.

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birth-scensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Data from Microsoft Bookshelf.

King, Bill."Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Notes

1. "Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).
2. "Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).
3. Data from West Magazine.
4. "CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).
5. "Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).
6. Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker
7. Data from West Magazine
8. "Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at https://web.archive.org/web/20130917120224/https://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).
9. "Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at https://web.archive.org/web/20130917120224/https://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).
10. "Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).
11. "Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).
12. "Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).
13. "Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Chapter 3 Extra Practice

3.1 Introduction to Bivariate Data

1. The following figure shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	The coastline	Near lakes and streams	One mountain peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Figure 3.26

a. Complete the table.

b. Are the events "being female" and "preferring the coastline" independent events?Let F = being female, and let C = preferring the coastline.

   - Find P(F AND C).
   - Find P(F)P(C)

   Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

c. Find the probability that a person is male, given that the person prefers hiking near lakes and streams. Let M = being male, and let L = preferring hiking near lakes and streams.

   - What word tells you this is a conditional?
   - Fill in the blanks, and calculate the probability: P(____|____) = ____.
   - Is the sample space for this problem all 100 hikers? If not, what is it?

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = preferring mountain peaks.

   - Find P(F).
   - Find P(P).
   - Find P(F AND P).
   - Find P(F OR P).

2. The figure below shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake path	Hilly path	Wooded path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

Figure 3.27

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

---

3. Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the Cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$, and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors, so the probability of choosing each door is $\frac{1}{3}$.

Caught or not	Door 1	Door 2	Door 3	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	
Not caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	
Total				1

Figure 3.28

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is P(Door 1 AND Caught)
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is P(Door 1 AND Not Caught)

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.
b. What is the probability that Alissa does not catch Muddy?
c. What is the probability that Muddy chooses Door 1 OR Door 2, given that Muddy is caught by Alissa?

4. The figure below relates the weights and heights of a group of individuals participating in an observational study.

Weight/height	Tall	Medium	Short	Total
Obsese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Total				

Figure 3.29

a. Find the total for each row and column.
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall, given that the individual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese, given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

---

5. There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Use the following information to answer the next four exercises. The figure below shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in school	Private instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Figure 3.30

a. Find *P*(musician is a female).
b. Find *P*(musician is a male AND had private instruction).
c. Find *P*(musician is a female OR is self-taught).
d. Are the events “being a female musician” and “learning music in school” mutually exclusive events?

6. An article in the *New England Journal of Medicine* reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.[1]

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking level	African American	Native Hawaiian	Latino	Japanese Americans	White	Total
1–10						
11–20						
21–30						
31+						
Total						

Figure 3.31

a. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.
b. Find the probability that the person was Latino.
c. In words, explain what it means to pick one person from the study who is "Japanese American **AND** smokes 21 to 30 cigarettes per day." Then find the probability.
d. In words, explain what it means to pick one person from the study who is "Japanese American **OR** smokes 21 to 30 cigarettes per day." Then find the probability.
e. In words, explain what it means to pick one person from the study who is "Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day." Then find the probability.
f. Prove that smoking level and ethnicity are dependent events.

7. The figure below contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the US.[2]

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	

Year	Robbery	Burglary	Rape	Vehicle	Total
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

Figure 3.32

Total each column and each row. Total data = 4,520.7.

a. Find *P* (2009 AND Robbery).
b. Find *P* (2010 AND Burglary).
c. Find *P* (2010 OR Burglary).
d. Find *P* (2011|Rape).
e. Find *P* (Vehicle|2008).

---

8. In an urn, there are 11 balls. Three balls are red (*R*), and eight balls are blue (*B*). Draw two balls, one at a time, with replacement. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows is seen in the figure below.

Total = 64 + 24 + 24 + 9 = 121

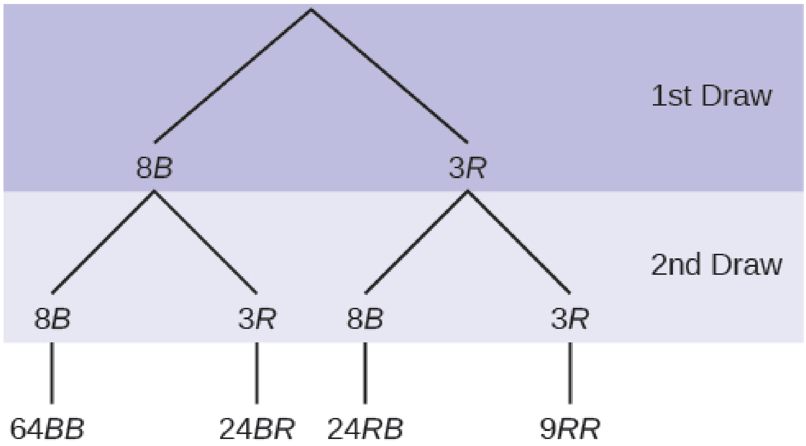


Figure 3.33. Figure description available at the end of the section.

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8. Then the nine *RR* outcomes can be written as:

R1R1R1R2R1R3R2R1R2R2R2R3R3R1R3R2R3R3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the sample space.

a. List the 24 BR outcomes: B1R1, B1R2, B1R3, …
b. Using the tree diagram, calculate P(RR).
c. Using the tree diagram, calculate P(RB OR BR).
d. Using the tree diagram, calculate P(R on first draw AND B on second draw).
e. Using the tree diagram, calculate P(R on second draw, GIVEN B on first draw).
f. Using the tree diagram, calculate P(BB).
g. Using the tree diagram, calculate P(B on the second draw, given R on the first draw).

---

9. An urn contains three red marbles and eight blue marbles. Draw two marbles, one at a time, from the urn, this time without replacement. "Without replacement" means that you do not put the first ball back before you select the second marble. Figure 3.34 shows a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example:

$$\left(\frac{3}{11}\right)\left(\frac{2}{10}\right)=\frac{6}{110}$$

Total = $\frac{56+24+24+6}{110}=\frac{110}{110}=1$

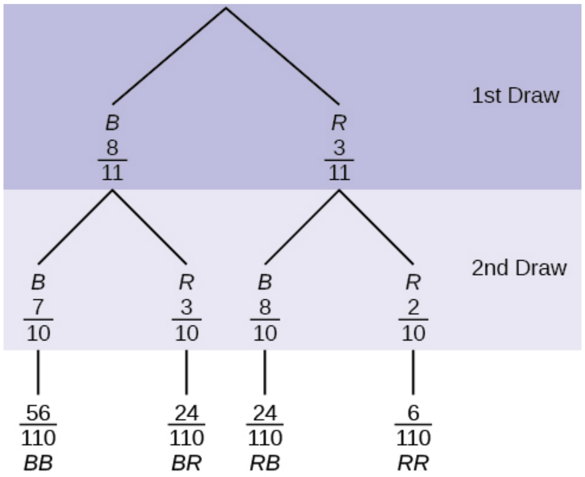


Figure 3.34. Figure description available at the end of the section.

NOTE: If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw without replacement, meaning there are ten marbles left in the urn on the second draw.

Calculate the following probabilities using the tree diagram.

a. P(RR)
b. P(RB OR BR) = $\left(\frac{3}{11}\right)\left(\frac{8}{10}\right)$ + (_____)(______)= $\frac{48}{110}$
c. P(R on second draw|B on first draw)
d. P(R on first draw AND B on second draw) = P(RB) = (_____)(______) = $\frac{24}{100}$
e. P(BB)
f. P(B on second draw|R on first draw)

---

10. In a standard deck, there are 52 cards. 12 cards are face cards (event F), and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate P(FF).

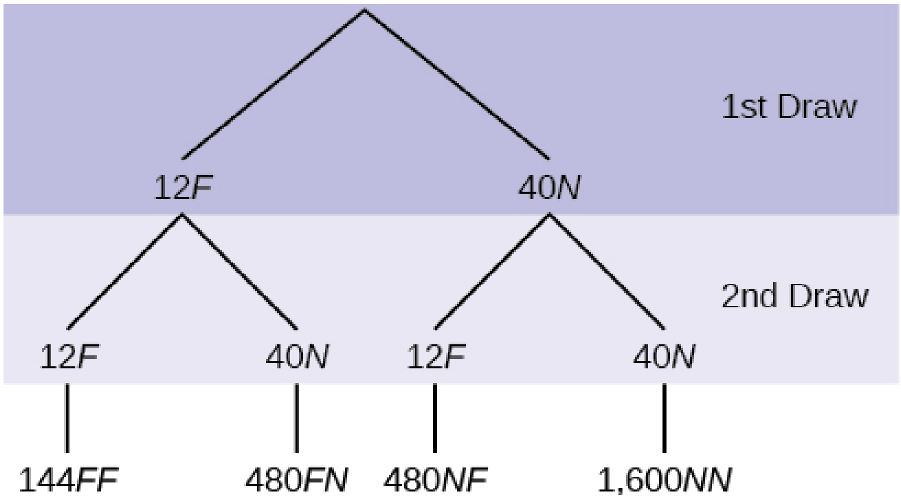


Figure 3.35. *Figure description available at the end of the section.*

11. In a standard deck, there are 52 cards. Twelve cards are face cards (F), and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram below is labeled with all possible probabilities.

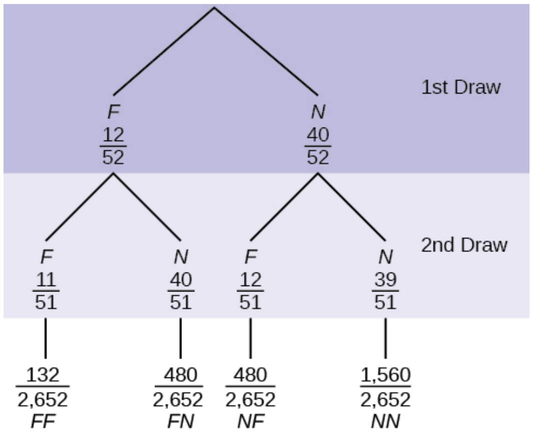


Figure 3.36. *Figure description available at the end of the section*.

a. Find P(FN OR NF).
b. Find P(N|F).
c. Find P(at most one face card).
Hint: "At most one face card" means zero or one face card.
d. Find P(at least one face card).
Hint: "At least one face card" means one or two face cards.

12. A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.

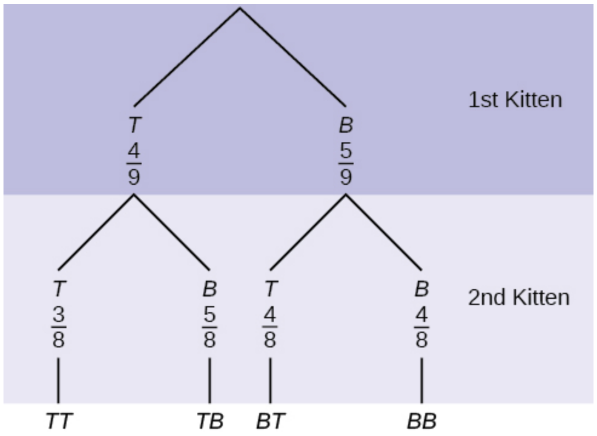


Figure 3.37. *Figure description available at the end of the section*.

1. What is the probability that both kittens are tabby?

a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$

b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$

c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$

d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$

2. What is the probability that one kitten of each coloring is selected?

a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$

b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$

c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$

d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$

3. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?

4. What is the probability of choosing two kittens of the same color?

---

13. Suppose there are four red balls and three yellow balls in a box. Two balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

---

14. Flip two fair coins. Let A = tails on the first coin and B = tails on the second coin. Then, A = {TT, *TH*} and B = {TT, *HT*}. Therefore:

- A AND B = {TT}
- A OR B = {*TH*, TT, *HT*}

The sample space when you flip two fair coins is X = {*HH*, *HT*, *TH*, *TT*}. The outcome *HH* is in NEITHER A NOR B. Draw a Venn diagram.

---

15. Roll a fair, six-sided die. Let A = a prime number of dots being rolled. Let B = an odd number of dots being rolled. If A = {2, 3, 5} and B = {1, 3, 5}, then:

- A AND B = {3, 5}
- A OR B = {1, 2, 3, 5}

The sample space for rolling a fair die is S = {1, 2, 3, 4, 5, 6}. Draw a Venn diagram representing this situation.

---

16. Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event C = {green, blue, purple} and event P = {red, yellow, blue}. Then C AND P = {blue} and C OR P = {green, blue, purple, red, yellow}. Draw a Venn diagram representing this situation.

---

17. Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, and 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let W = working a second job and S = spouse also working.

---

18. A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative Rh factor, 5–10% of African Americans have the Rh- factor, and 51% have type O blood.[3] [4]

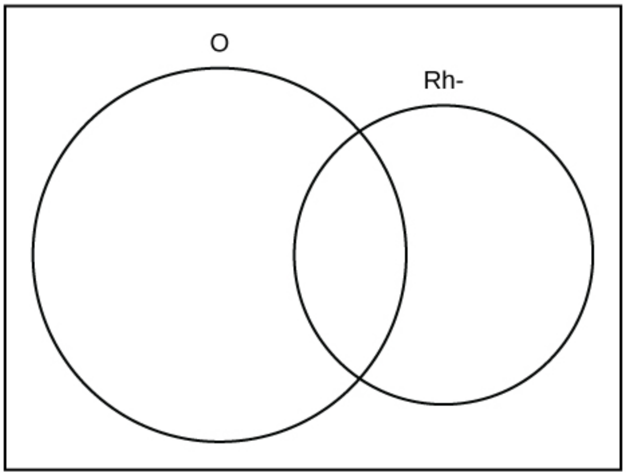


Figure 3.38. Figure description available at the end of the section.

The “O” circle represents the African Americans with type O blood. The “Rh-” oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10%, using 7.5% as the percent of African Americans who have the Rh- factor. Let O = African American with type O blood and R = African American with Rh- factor.

a. Find P(O).
b. Find P(R).
c. Find P(O AND R).
d. Find P(O OR R).
e. In a complete sentence, describe the overlapping area of the Venn diagram.
f. In a complete sentence, describe the area of the Venn diagram in the rectangle but outside both the circle and the oval.

---

19. In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or a non-fiction book.
c. In the Venn diagram, describe the overlapping area using a complete sentence.
d. Suppose that some customers buy only compact discs. Draw an oval in your Venn diagram representing this event.

---

20. The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51.[5] Let: C = a man developing cancer in his lifetime and P = man having at least one false positive. Construct a tree diagram of the situation.

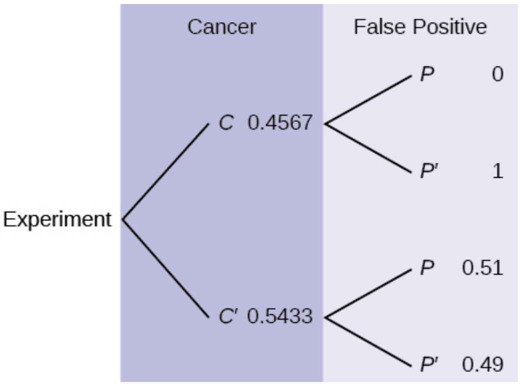


Figure 3.39. Figure description available at the end of the section.

21. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (*R*), four yellow (*Y*), and five blue (*B*) beads. For the coin, *P*(*H*) = $\frac{2}{3}$ and *P*(*T*) = $\frac{1}{3}$, where *H* is heads and *T* is tails.

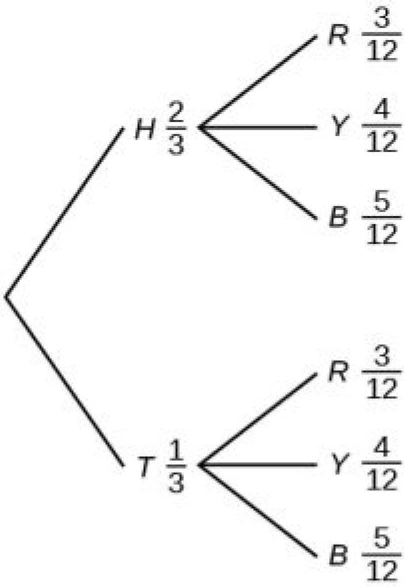


Figure 3.40. Figure description available at the end of the section.

Find *P*(tossing a head on the coin AND a red bead).

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

Find *P*(blue bead).

a. $\frac{15}{36}$
b. $\frac{10}{36}$
c. $\frac{10}{12}$
d. $\frac{6}{36}$

22. A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it.

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the *second* cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let S be the event that both cookies selected were the same flavor. Find P(S).
e. Let T be the event that the cookies selected were different flavors. Find P(T) by two different methods, using the complement rule and using the branches of the tree. Your answers should be the same with both methods.
f. Let U be the event that the second cookie selected is a butter cookie. Find P(U).

---

23. Suppose that you have eight cards. Five are green, and three are yellow. The cards are well shuffled.

Suppose that you randomly draw two cards, one at a time, with replacement. Let G_1 = first card is green and G_2 = second card is green.

a. Draw a tree diagram of the situation.
b. Find P(G_1 AND G_2).
c. Find P(at least one green).
d. Find P($G_2|G_1$).
e. Are G_2 and G_1 independent events? Explain why or why not.

Suppose that you randomly draw two cards, one at a time, without replacement. Let G_1 = first card is green and G_2 = second card is green.

a. Draw a tree diagram of the situation.
b. Find P(G_1 AND G_2).
c. Find P(at least one green).
d. Find P($G_2|G_1$).
e. Are G_2 and G_1 independent events? Explain why or why not.

---

24. The percent of licensed US drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; and 13.61% are age 65 or over. Of the licensed US male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; and 13.53% are age 65 or over.[6]

Complete the following:

a. Construct a table or a tree diagram of the situation.
b. Find P(driver is female).
c. Find P(driver is age 65 or over|driver is female).
d. Find P(driver is age 65 or over AND female).
e. In words, explain the difference between the probabilities in part (c) and part (d).
f. Find P(driver is age 65 or over).
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

Suppose that 10,000 US licensed drivers are randomly selected.

a. How many would you expect to be male?
b. Using the table or tree diagram, construct a contingency table of gender versus age group.
c. Using the contingency table, find the probability that a driver randomly selected from the 20–64 age group is female.

---

25. Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work, and approximately 5.3% take public transportation.[7]

a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

---

26. When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that 140 of 250 spins showed a head (event H), while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

a. Based on the given data, find $P(H)$ and $P(T)$.
b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
d. Use the tree to find the probability of obtaining at least one head.

27. The following are real data from Santa Clara County, CA. At a certain point in time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:[8]

	Homosexual/ bisexual	IV drug user*	Heterosexual contact	Other	Total
Female	0	70	136	49	
Male	2,146	463	60	135	
Total					
*includes homosexual/bisexual IV drug users					

Figure 3.41

Suppose a person with AIDS in Santa Clara County is randomly selected.

a. Find *P*(person is female).
b. Find *P*(person has a risk factor heterosexual contact).
c. Find *P*(person is female OR has a risk factor of IV drug user).
d. Find *P*(person is female AND has a risk factor of homosexual/bisexual).
e. Find *P*(person is male AND has a risk factor of IV drug user).
f. Find *P*(person is female, GIVEN person got the disease from heterosexual contact).
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

The completed contingency table is as follows:

	Homosexual/ bisexual	IV drug user*	Heterosexual contact	Other	Total
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Total	**2,146**	**533**	**196**	**184**	**3,059**
*includes homosexual/bisexual IV drug users					

Figure 3.42

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

a. Find *P*(person is female).
b. Find *P*(person obtained the disease through heterosexual contact).
c. Find *P*(person is female, GIVEN person got the disease from heterosexual contact).
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

28. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012 and when they are up for reelection.[9]

Up for reelection	Democratic party	Republican party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Figure 3.43

a. What is the probability that a randomly selected senator has an "Other" affiliation?
b. What is the probability that a randomly selected senator is up for reelection in November 2016?
c. What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?
d. What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?
e. Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?
f. Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?
g. The events "Republican" and "Up for reelection in 2016" are ________.

    - mutually exclusive
    - independent
    - both mutually exclusive and independent
    - neither mutually exclusive nor independent

h. The events "Other" and "Up for reelection in November 2016" are ________.

    - mutually exclusive
    - independent
    - both mutually exclusive and independent
    - neither mutually exclusive nor independent

---

29. The figure below gives the number of suicides estimated in the US for a recent year by age, race (Black or White), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population.

Race and sex	1-14	15-24	25-64	Over 64	Total
White, male	210	3,360	13,610		22,050
White, female	80	580	3,380		4,930
Black, male	10	460	1,060		1,670

Race and sex	1-14	15-24	25-64	Over 64	Total
Black, female	0	40	270		330
All others					
Total	310	4,650	18,780		29,760

Figure 3.44

Do not include "all others."

a. Fill in the column for the suicides for individuals over age 64.
b. Fill in the row for all other races.
c. Find the probability that a randomly selected individual was a White male.
d. Find the probability that a randomly selected individual was a Black female.
e. Find the probability that a randomly selected individual was Black.
f. Find the probability that a randomly selected individual was a Black or White male. Do not include "all others."
g. Out of the individuals over age 64, find the probability that a randomly selected individual was a Black or White male. Do not include "all others."

---

30. The table of data obtained from the website Baseball Almanac shows hit information for four well-known baseball players. Suppose that one hit from the table is randomly selected.[10]

Name	Single	Double	Triple	Home run	Total hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Figure 3.45

Find P(hit was made by Babe Ruth).

a. $\frac{1{,}518}{2{,}873}$
b. $\frac{2{,}873}{12{,}351}$
c. $\frac{583}{12{,}351}$
d. $\frac{4{,}189}{12{,}351}$

Find P(hit was made by Ty Cobb|hit was a home run).

a. $\frac{4,189}{12,351}$
b. $\frac{114}{1,720}$
c. $\frac{1,720}{4,189}$
d. $\frac{114}{12,351}$

31. The figure below identifies a group of children by one of four hair colors and by type of hair.

Hair type	Brown	Blond	Black	Red	Total
Wavy	20		15	3	43
Straight	80	15		12	
Total		20			215

Figure 3.46

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If B is the event of a child having brown hair, find the probability of the complement of B.
g. In words, what does the complement of B represent?

32. In a previous year, the weights of the members of the San Francisco 49ers and the Dallas Cowboys were published in the *San Jose Mercury News*. The factual data were compiled into the following table.[11]

Shirt number	≤ 210	211–250	251–290	> 290
1-33	21	5	0	0
34-66	6	18	7	4
66-99	6	12	22	5

Figure 3.47

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33, GIVEN that he weighs at most 210 pounds.

3.2 Visualizing Bivariate Quantitative Data

1. The Gross Domestic Product Purchasing Power Parity (GDP PPP) is an indication of a country's currency value compared to another country. The figure below shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Figure 3.48

2. The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data.

Year	Poverty rate	Cellular usage per capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Figure 3.49

3. Does higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data. Note that tuition is the independent variable and salary is the dependent variable.

School	Mid-career salary (in thousands)	Yearly tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050

School	Mid-career salary (in thousands)	Yearly tuition
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Figure 3.50

3.3 Measures of Association

1. Can a coefficient of determination be negative? Why or why not?

2. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. The figure below shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Figure 3.51

Find r and r^2.

3. The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data.

Year	Poverty rate	Cellular usage per capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86

Year	Poverty rate	Cellular usage per capita
2009	12	90.82

Figure 3.52

Find r and r^2.

4. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data. Note that tuition is the independent variable and salary is the dependent variable.

School	Mid-career salary (in thousands)	Yearly tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Figure 3.53

Find r and r^2.

3.4 Modeling Linear Relationships

1. A random sample of ten professional athletes produced the following data, where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Figure 3.54

a. Draw a scatter plot of the data.
b. Use regression to find the equation for the line of best fit.
c. Draw the line of best fit on the scatter plot.
d. What is the slope of the line of best fit? What does it represent?
e. What is the y-intercept of the line of best fit? What does it represent?
f. What does an r value of zero mean?
g. When n = 2 and r = 1, are the data significant? Explain.
h. When n = 100 and r = -0.89, is there a significant correlation? Explain.

---

2. What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

3.5 Cautions about Regression

1. The following table shows economic development measured in per capita income (PCINC).

Year	PCINC	year	PCINC
1870	340	1920	1050

Year	PCINC	year	PCINC
1880	499	1930	1170
1890	592	1940	1364
1900	757	1950	1836
1910	927	1960	2132

Figure 3.55

a. What are the independent and dependent variables?
b. Draw a scatter plot.
c. Use regression to find the line of best fit and the correlation coefficient.
d. Interpret the significance of the correlation coefficient.
e. Is there a linear relationship between the variables?
f. Find the coefficient of determination and interpret it.
g. What is the slope of the regression equation? What does it mean?
h. Use the line of best fit to estimate PCINC for 1900 and for 2000.
i. Determine if there are any outliers.

---

2. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

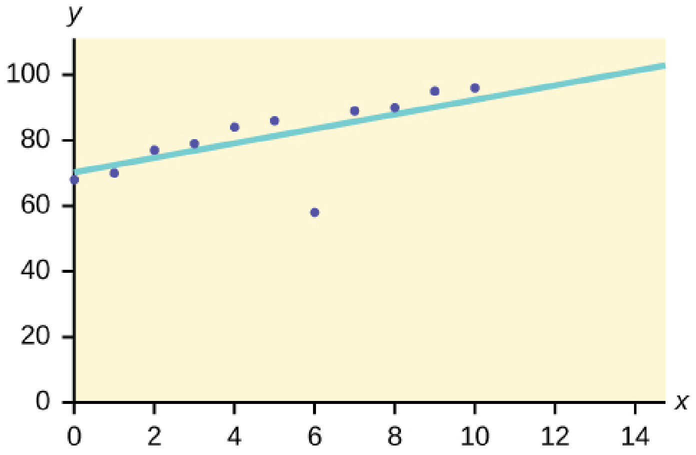


Figure 3.56. Figure description available at the end of the section.

a. Do there appear to be any outliers?
b. A point is removed, and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?
c. What effect did the potential outlier have on the line of best fit?
d. Are you more or less confident in the predictive ability of the new line of best fit?

e. The sum of squared errors for a dataset of 18 numbers is 49. What is the standard deviation?
f. The standard deviation for the sum of squared errors for a dataset is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

---

3. The heights (sidewalk to roof) of notable tall buildings in America are compared to the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

Figure 3.57

a. Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Based on the data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least-squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least-squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

---

4. Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their popu-

lation. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.

Percent returning: 74, 66, 81, 52, 73, 62, 52, 45, 62, 46, 60, 46, 38
Percent new: 5, 6, 8, 11, 12, 15, 16, 17, 18, 18, 19, 20, 20

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from (a).
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?

5. The following table shows data on average per capita coffee consumption and heart disease rate in a random sample of ten countries.

Coffee consumption and heart disease										
Yearly coffee consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

Figure 3.58

a. Enter the data into your calculator, and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from (a).
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. Do the data provide convincing evidence that there is a linear relationship between the amount of coffee consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

6. The following table consists of one student athlete's time (in minutes) to swim 2,000 yards and the student's heart rate (beats per minute) after swimming on a random sample of ten days:

Swim time	Heart rate
34.12	144

Swim time	Heart rate
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

Figure 3.59

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from (a).
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

---

7. A researcher is investigating whether population impacts homicide rate. He uses demographic data from Detroit, MI, to compare homicide rates and the number of the population that are White males.

Population size	Homicide rate per 100,000 people
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

Figure 3.60

a. Use your calculator to construct a scatter plot of the data. What should the independent variable be? Why?
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
c. Discuss what the following mean in context:
   i. The slope of the regression equation
   ii. The y-intercept of the regression equation
   iii. The correlation r
   iv. The coefficient of determination r^2
d. Do the data provide convincing evidence that there is a linear relationship between population size and homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

---

8. Use the table below to answer (a) and (b).

School	Mid-career salary (in thousands)	Yearly tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Figure 3.61

1. Use the data to determine the linear-regression line equation with the outliers removed. Is there a linear correlation for the dataset with outliers removed? Justify your answer.
2. If we remove the two service academies (the tuition is \$0.00), we construct a new regression equation of $y = -0.0009x + 160$ with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the dataset.

---

9. The average number of people in a family that attended college for various years is given below.

Year	Number of family members attending college
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

Figure 3.62

a. Using "Year" as the independent variable and "Number of Family Members Attending College" as the dependent variable, draw a scatter plot of the data.
b. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
c. Find the correlation coefficient. Is it significant?
d. Pick two years between 1969 and 1991, and find the estimated number of family members attending college.
e. Based on the data, is there a linear relationship between the year and the average number of family members attending college?
f. Using the least-squares line, estimate the number of family members attending college for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
g. Are there any outliers in the data?
h. What is the estimated average number of family members attending college for 1986? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

---

10. The percent of female wage and salary workers who are paid hourly rates is given in below for the years 1979 to 1992.

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7

Year	Percent of workers paid hourly rates
1990	62.8
1992	62.9

Figure 3.63

a. Using the year as the independent variable and the percent as the dependent variable, draw a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated percents for 1988 and 1991.
f. Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
g. Are there any outliers in the data?
h. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

---

11. The cost of a leading liquid laundry detergent in different sizes is given below.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.00	
200	10.99	

Figure 3.64

a. Complete the table for the cost per ounce of the different sizes.
b. Using size as the independent variable and cost per ounce as the dependent variable, draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. If the laundry detergent were sold in a 40-ounce size, find the estimated cost per ounce.
g. If the laundry detergent were sold in a 90-ounce size, find the estimated cost per ounce.
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the data?
j. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would cost per ounce? Why or why not?

k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

---

12. According to a flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net taxable estate ($)	Approximate probate fees and taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

Figure 3.65

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated total cost for a next taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

13. The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale price ($)
9	147
20	197
27	297

Size (inches)	Sale price ($)
31	447
35	1,277
40	2,177
60	2,497

Figure 3.66

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated sale price for a 32-inch television. Find the cost for a 50-inch television.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

14. The figure below shows the average heights for American boys in 1990.

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

Figure 3.67

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated average height for a one-year-old. Find the estimated average height for an 11-year-old.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?

i. Use the least-squares line to estimate the average height for a 62-year-old man. Do you think that your answer is reasonable? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

---

15. Use the table below to answer (a)–(n).

State	Number of letters in name	Year entered the Union	Ranks for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Figure 3.68

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

a. What are the independent and dependent variables?
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in (f) to plot the least-squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least-squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
k. Delete "Hawaii" and substitute "Alaska" for it. Alaska is the 49th state, with an area of 656,424 square miles. Calculate the new least-squares line.
l. Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that's the case?
m. Do you think that, in general, newer states are larger than the original states?

Figure Descriptions

Figure 3.33: This is a tree diagram with branches showing frequencies of each draw. The first branch shows two lines: 8B and 3R. The second branch has a set of two lines (8B and 3R) for each line of the first branch. Multiply along each line to find 64BB, 24BR, 24RB, and 9RR.

Figure 3.34: This is a tree diagram with branches showing probabilities of each draw. The first branch shows two lines: B 8/11 and R 3/11. The second branch has a set of two lines for each first branch line. Below B 8/11 are B 7/10 and R 3/10. Below R 3/11 are B 8/10 and R 2/10. Multiply along each line to find BB 56/110, BR 24/110, RB 24/110, and RR 6/110.

Figure 3.35: This is a tree diagram with branches showing frequencies of each draw. The first branch shows two lines: 12F and 40N. The second branch has a set of two lines (12F and 40N) for each line of the first branch. Multiply along each line to find 144FF, 480FN, 480NF, and 1,600NN.

Figure 3.36: This is a tree diagram with branches showing frequencies of each draw. The first branch shows two lines: F 12/52 and N 40/52. The second branch has a set of two lines (F 11/52 and N 40/51) for each line of the first branch. Multiply along each line to find FF 121/2652, FN 480/2652, NF 480/2652, and NN 1560/2652.

Figure 3.37: This is a tree diagram with branches showing probabilities of kitten choices. The first branch shows two lines: T 4/9 and B 5/9. The second branch has a set of two lines for each first branch line. Below T 4/9 are T 3/8 and B 5/8. Below B 5/9 are T 4/8 and B 4/8. Multiply along each line to find probabilities of possible combinations.

Figure 3.38: This is an empty Venn diagram showing two overlapping circles. The left circle is labeled O and the right circle is labeled RH-.

Figure 3.39: This is a tree diagram with two branches. The first branch, labeled Cancer, shows two lines: 0.4567 C and 0.5433 C'. The second branch is labeled False Positive. From C, there are two lines: zero P and one P'. From C', there are two lines: 0.51 P and 0.49 P'.

Figure 3.40: Tree diagram with two branches. The first branch consists of two lines of H=2/3 and T=1/3. The second branch consists of two sets of three lines each with the both sets containing R=3/12, Y=4/12, and B=5/12.

Figure 3.56: Scatterplot that shows a positive, very linear pattern of dots except for one (6, 60). A line of best fit is overlayed on the graph that closely follows the linear dots.

References

Figures

Figure 3.33: Figure 3.10 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams

Figure 3.34: Figure 3.13 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams

Figure 3.35: Figure 3.12 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams

Figure 3.36: Figure 3.14 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams

Figure 3.37: Figure from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams

Figure 3.38: Figure 3.18 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams

Figure 3.39: Figure 3.15 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/3-solutions#element-750-solution

Figure 3.40: Figure 3.22 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/3-homework

Figure 3.56: Figure 12.24 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/12-practice

Text

Data from the House Ways and Means Committee, the Health and Human Services Department.

Data from Microsoft Bookshelf.

Data from the United States Department of Labor, the Bureau of Labor Statistics.

Data from the Physician's Handbook, 1990.

Notes

1. Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." *The New England Journal of Medicine*, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).
2. "United States: Uniform Crime Report – State Statistics from 1960–2011." The Disaster Center. Available online at http://www.disastercenter.com/crime (accessed May 2, 2013).
3. "Blood Types." American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-about-blood/bloodtypes (accessed May 3, 2013).
4. Samuel, T. M. "Strange Facts about RH Negative Blood." Healthfully, 2017. Available online at https://healthfully.com/strange-rh-negative-blood-5552003.html (accessed January 26, 2021).
5. Data from the American Cancer Society.

6. Data from the Federal Highway Administration, part of the United States Department of Transportation.
7. Data from the Federal Highway Administration, part of the United States Department of Transportation.
8. Data from Santa Clara County Public Health Department
9. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).
10. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).
11. Data from the San Jose Mercury News.

Chapter 4 Extra Practice

4.1 Introduction to Probability and Random Variables

1. You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), Q (queen), and *K* (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are Q of spades, *K* of hearts, and Q of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are Q of spades, *K* of hearts, and *J* of spades. Can you decide if the sampling was with or without replacement?

---

2. You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), Q (queen), and *K* (king) of that suit. Let S = spades, *H* = hearts, *D* = diamonds, and C = clubs.

a. Suppose you pick four cards but do not put any cards back into the deck. Your cards are QS, 1*D*, 1C, and Q*D*.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are *KH*, 7*D*, 6*D*, and *KH*.

Of (a) and (b), which did you sample with replacement, and which did you sample without replacement?

---

3. You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), Q (queen), and *K* (king) of that suit. Let S = spades, *H* = hearts, *D* = diamonds, and *C* = clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

QS, 1D, 1C, QD

KH, 7D, 6D, KH

QS, 7D, 6D, KS

4. Flip two fair coins. The sample space is {*HH*, *HT*, *TH*, *TT*}, where *T* = tails and *H* = heads. The outcomes are *HH*, *HT*, *TH*, and *TT*. The outcomes *HT* and *TH* are different. The *HT* means that the first coin showed heads and the second coin showed tails. The *TH* means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting at most one tail (meaning zero or one tail). Then A can be written as {*HH*, *HT*, *TH*}. The outcome *HH* shows zero tails. *HT* and *TH* each show one tail.
- Let B = the event of getting all tails. B can be written as {*TT*}. B is the complement of A, so B = A′. In addition, *P*(A) + *P*(B) = *P*(A) + *P*(A′) = 1.
- The probabilities for A and for B are *P*(A) = $\frac{3}{4}$ and *P*(B) = $\frac{1}{4}$.
- Let C = the event of getting all heads. C = {*HH*}. Since B = {*TT*}, *P*(B AND C) = 0. B and C are mutually exclusive, with no members in common because you cannot have all tails and all heads at the same time.
- Let *D* = event of getting more than one tail. *D* = {*TT*}. *P*(*D*) = $\frac{1}{4}$.
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) E = {*HT*, *HH*}. *P*(*E*) = $\frac{2}{4}$.
- Find the probability of getting at least one (one or two) tail in two flips. Let *F* = event of getting at least one tail in two flips. *F* = {*HT*, *TH*, *TT*}. *P*(*F*) = $\frac{3}{4}$.

5. Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

6. Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd, so A = {1, 3, 5}. Let event B = a face is even, so B = {2, 4, 6}.

a. Find the complement of A, A′. The complement of A, A′ is B because A and B together make up the sample space. P(A) + P(B) = P(A) + P(A′) = 1. Additionally, P(A) = $\frac{3}{6}$ and P(B) = $\frac{3}{6}$.
b. Let event C = odd faces larger than two, so C = {3, 5}. Let event *D* = all even faces smaller than five, so *D* = {2, 4}. P(C AND *D*) = 0 because you cannot have an odd and even face at the same time. Therefore, C and *D* are mutually exclusive events.
c. Let event E = all faces less than five, so E = {1, 2, 3, 4}.
d. Find P(C|A). This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find P(C|A), find the probability of C using the sample space, A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, *P*(C|A) = $\frac{2}{3}$.

Are C and E mutually exclusive events? Why or why not?

7. Let event A = learning Spanish. Let event B = learning German. Then, A AND B = learning Spanish and German. Suppose P(A) = 0.4 and P(B) = 0.2. P(A AND B) = 0.08. Are events A and B independent? Hint: You must show ONE of the following:

- P(A|B) = P(A)
- P(B|A) = P(B)
- P(A AND B) = P(A)P(B)

---

8. Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose P(G) = 0.6, P(H) = 0.5, and P(G AND H) = 0.3. Are G and H independent?

If G and H are independent, then you must show ONE of the following:

- P(G|H) = P(G)
- P(H|G) = P(H)
- P(G AND H) = P(G)P(H)

NOTE: The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

---

9. In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is S = {R1, R2, R3, R4, R5, R6, G1, G2, G3, G4}.

S has ten outcomes. What is P(G AND O)?

---

10. A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that P(B) = 0.40, P(D) = 0.30, and P(B AND D) = 0.20.

a. Find P(B|D).
b. Find P(D|B).
c. Are B and D independent?
d. Are B and D mutually exclusive?

11. In a box, there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well shuffled. You reach into the box (you cannot see into it) and draw one card.

Let R = red card is drawn, B = blue card is drawn, and E = even-numbered card is drawn.

The sample space S = R1, R2, R3, B1, B2, B3, B4, B5. S has eight outcomes.

- P(R) = $\frac{3}{8}$. P(B) = $\frac{5}{8}$. P(R AND B) = 0 (since you cannot draw one card that is both red and blue).
- P(E) = $\frac{3}{8}$. (There are three even-numbered cards: R2, B2, and B4.)
- P(E|B) = $\frac{2}{5}$. (There are five blue cards: B1, B2, B3, B4, and B5. Out of the blue cards, there are two even cards: B2 and B4.)
- P(B|E) = $\frac{2}{3}$. (There are three even-numbered cards: R2, B2, and B4. Out of the even-numbered cards, two are blue: B2 and B4.)
- The events R and B are mutually exclusive because P(R AND B) = 0.
- Let G = card with a number greater than 3. G = {B4, B5}. P(G) = $\frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. H = {B1, B2, B3, B4}. P(G|H) = $\frac{1}{4}$. (The only card in H that has a number greater than three is B4.) Since $\frac{2}{8}$ = $\frac{1}{4}$, P(G) = P(G|H), which means that G and H are independent.

---

12. In a basketball arena:

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.

Let B be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

---

13. In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F = the event that a student is female. Let L = the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

The following probabilities are given in this example:

- P(F) = 0.60; P(L) = 0.50
- P(F AND L) = 0.45
- P(L|F) = 0.75

NOTE: The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know P(F|L) yet, so you cannot use the second condition.

---

14. Mark is deciding which route to take to work. His choices are the interstate (I) and Fifth Street (F).

- P(I) = 0.44 and P(F) = 0.56
- P(I AND F) = 0 because Mark will take only one route to work.

What is the probability of P(I OR F)?

---

15. Fill in the blanks to the following questions.

a. Toss one fair coin with two sides, H and T. The outcomes are _______. How many outcomes are there?
b. Toss one fair, six-sided die (with 1, 2, 3, 4, 5, or 6 dots on a side). The outcomes are _______. How many outcomes are there?
c. Multiply the two numbers of outcomes. The answer is _______.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer from (c) is the number of outcomes (i.e., the size of the sample space). What are the outcomes? (Hint: Two of the outcomes are H1 and T6.)
e. Event A = heads (H) on the coin followed by an even number (2, 4, 6) on the die.
   A = {_______}. Find P(A).
f. Event B = heads on the coin followed by a three on the die. B = {_______}. Find P(B).
g. Are A and B mutually exclusive? (Hint: What is P(A AND B)? If P(A AND B) = 0, then A and B are mutually exclusive.)
h. Are A and B independent? (Hint: Does P(A AND B) = P(A)P(B)? If P(A AND B) = P(A)P(B), then A and B are independent. If not, then they are dependent.)

---

16. A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let T be the event of getting the white ball twice, F the event of picking the white ball first, and S the event of picking the white ball in the second drawing.

a. Compute P(T).
b. Compute P(T|F).
c. Are T and F independent?

d. Are F and S mutually exclusive?
e. Are F and S independent?

---

17. E and F are mutually exclusive events. P(E) = 0.4; P(F) = 0.5. Find P(E|F).

---

18. J and K are independent events. P(J|K) = 0.3. Find P(J).

---

19. U and V are mutually exclusive events. P(U) = 0.26; P(V) = 0.37. Find:

a. P(U AND V)
b. P(U|V)
c. P(U OR V)

---

20. Q and R are independent events. P(Q) = 0.4 and P(Q AND R) = 0.1. Find P(R).

---

21. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.

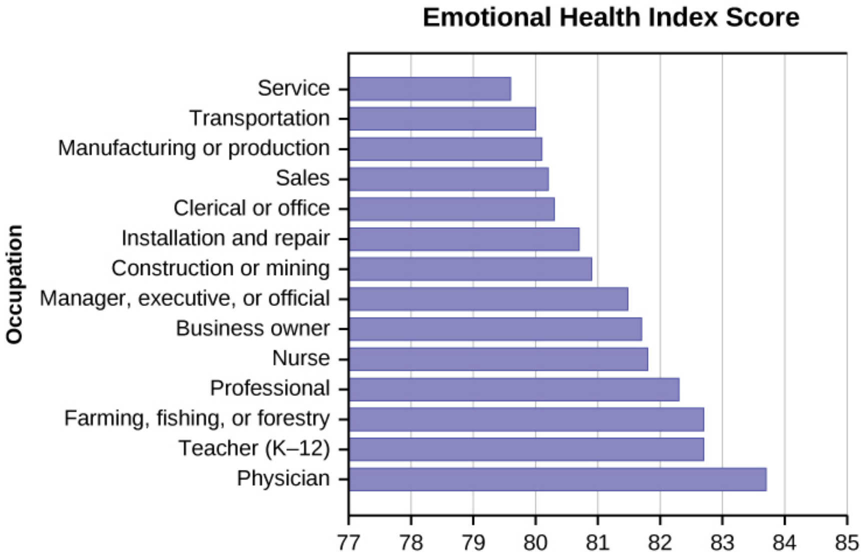


Figure 4.25. *Figure description available at the end of the section*.

a. Find the probability that an Emotional Health Index Score is 82.7.
b. Find the probability that an Emotional Health Index Score is 81.0.
c. Find the probability that an Emotional Health Index Score is more than 81.
d. Find the probability that an Emotional Health Index Score is between 80.5 and 82.
e. If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?
f. What is the probability that an Emotional Health Index Score is 80.7 or 82.7?
g. What is the probability that an Emotional Health Index Score is less than 80.2, given that it is already less than 81
h. What occupation has the highest Emotional Health Index Score?
i. What occupation has the lowest Emotional Health Index Score?
j. What is the range of the data?
k. Compute the average Emotional Health Index Score.
l. If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

---

22. A previous year, the weights of the members of the San Francisco 49ers and the Dallas Cowboys were published in the *San Jose Mercury News*. The factual data are compiled into the figure below.[1]

Shirt number	≤ 210	211–250	251–290	290≤
1-33	21	5	0	0
34-66	6	18	7	4
66-99	6	12	22	5

Figure 4.26

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about *P*(Shirt number 1–33|≤ 210 pounds)?

---

23. The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write “not enough information” for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

a. Find *P*(C).
b. Find *P*(*P*|C).
c. Find *P*(*P*|C’).
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify

numerically, and explain why or why not.

---

24. Given events G and H: P(G) = 0.43; P(H) = 0.26; P(H AND G) = 0.14

a. Find P(H OR G).
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

---

25. Given events J and K: P(J) = 0.18; P(K) = 0.37; P(J OR K) = 0.45

a. Find P(J AND K).
b. Find the probability of the complement of event (J AND K).
c. Find the probability of the complement of event (J OR K).

---

26. The sample space S is the whole numbers starting at one and less than 20.

a. Find S.
b. Let event A = the even numbers and event B = numbers greater than 13. Find A and B.
c. Find P(A) and P(B).
d. Find A AND B. Find A OR B.
e. Find P(A AND B). Find P(A OR B).
f. Find A′. Find P(A′).
g. What is P(A) + P(A′)?
h. Find P(A|B) and P(B|A). Are the probabilities equal?

---

27. The sample space S is all the ordered pairs of two whole numbers, the first from one to three and the second from one to four—for example, (1, 4).

a. Find S.
b. Let event A = the sum is even and event B = the first number is prime. Find A and B.
c. Find P(A) and P(B).
d. Find A AND B. Find A OR B.
e. Find P(A AND B). Find P(A OR B).
f. Find B′. Find P(B′).
g. What is P(A) + P(A′)?
h. Find P(A|B) and P(B|A). Are the probabilities equal?

28. A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S, and compute its probability. *NOTE*: an outcome is the number of dots that show up.

a. Event T = the outcome is two
b. Event A = the outcome is an even number
c. Event B = the outcome is less than four
d. The complement of A
e. A GIVEN B
f. B GIVEN A
g. A AND B
h. A OR B
i. A OR B′
j. Event N = the outcome is a prime number
k. Event I = the outcome is seven

29. The figure below describes the distribution of a random sample, S, of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Male	43	9
Female	44	4

Figure 4.27

Let the events M = the subject is male, F = the subject is female, R = the subject is right-handed, and L = the subject is left-handed. Compute the following probabilities:

a. P(M)
b. P(F)
c. P(R)
d. P(L)
e. P(M AND R)
f. P(F AND L)
g. P(M OR F)
h. P(M OR R)
i. P(F OR L)
j. P(M')
k. P(R|M)
l. P(F|L)
m. P(L|F)

30. In a particular college class, there are male and female students. Some students have long hair, and some students have short hair. Write the symbols for the probabilities of the events (a) through (j). (Note that you cannot find numerical answers here; you were not given enough information to find any probability values yet. Concentrate on understanding the symbols.)

- Let *F* = the event that a student is female.
- Let *M* = the event that a student is male.
- Let *S* = the event that a student has short hair.
- Let *L* = the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

---

31. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

- Let *H* = the event of getting a hat.
- Let *N* = the event of getting a noisemaker.
- Let *F* = the event of getting a finger trap.
- Let *C* = the event of getting a bag of confetti.

a. Find *P*(*H*).
b. Find *P*(*N*).
c. Find *P*(*F*).
d. Find *P*(*C*).

32. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

- Let *B* = the event of getting a blue jelly bean
- Let *G* = the event of getting a green jelly bean.
- Let *O* = the event of getting an orange jelly bean.

- Let P = the event of getting a purple jelly bean.
- Let R = the event of getting a red jelly bean.
- Let Y = the event of getting a yellow jelly bean.

1. Find P(B).
2. Find P(G).
3. Find P(P).
4. Find P(R).
5. Find P(Y).
6. Find P(O).

---

33. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

- Let A = the event that a country is in Asia.
- Let E = the event that a country is in Europe.
- Let F = the event that a country is in Africa.
- Let N = the event that a country is in North America.
- Let O = the event that a country is in Oceania.
- Let S = the event that a country is in South America.

a. Find P(A).
b. Find P(E).
c. Find P(F).
d. Find P(N).
e. Find P(O).
f. Find P(S).

---

34. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

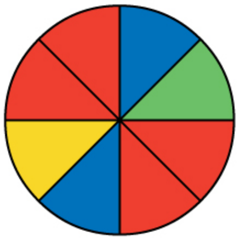


Figure 4.28. *Figure description available at the end of the section.*

- Let B = the event of landing on blue.
- Let R = the event of landing on red.
- Let G = the event of landing on green.
- Let Y = the event of landing on yellow.

a. If you land on Y, you get the biggest prize. Find $P(Y)$.
b. If you land on red, you don't get a prize. What is $P(R)$?

---

35. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

- Let I = the event that a player in an infielder.
- Let O = the event that a player is an outfielder.
- Let H = the event that a player is a great hitter.
- Let N = the event that a player is not a great hitter.

a. Write the symbols for the probability that a player is not an outfielder.
b. Write the symbols for the probability that a player is an outfielder or is a great hitter.
c. Write the symbols for the probability that a player is an infielder and is not a great hitter.
d. Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.
e. Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.
f. Write the symbols for the probability that, of all the outfielders, a player is not a great hitter.
g. Write the symbols for the probability that, of all the great hitters, a player is an outfielder.
h. Write the symbols for the probability that a player is an infielder or is not a great hitter.
i. Write the symbols for the probability that a player is an outfielder and is a great hitter.
j. Write the symbols for the probability that a player is an infielder.
k. What is the word for the set of all possible outcomes?
l. What is conditional probability?
m. The likelihood that an event will occur, given that another event has already occurred.

36. A shelf holds 12 books. Eight are fiction, and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book.

- Let F = event that book is fiction
- Let N = event that book is nonfiction

a. What is the sample space?
b. What is the sum of the probabilities of an event and its complement?

---

37. You are rolling a fair, six-sided die. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

a. What does P(E|M) mean in words?
b. What does P(E OR M) mean in words?

---

38. The graph below displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

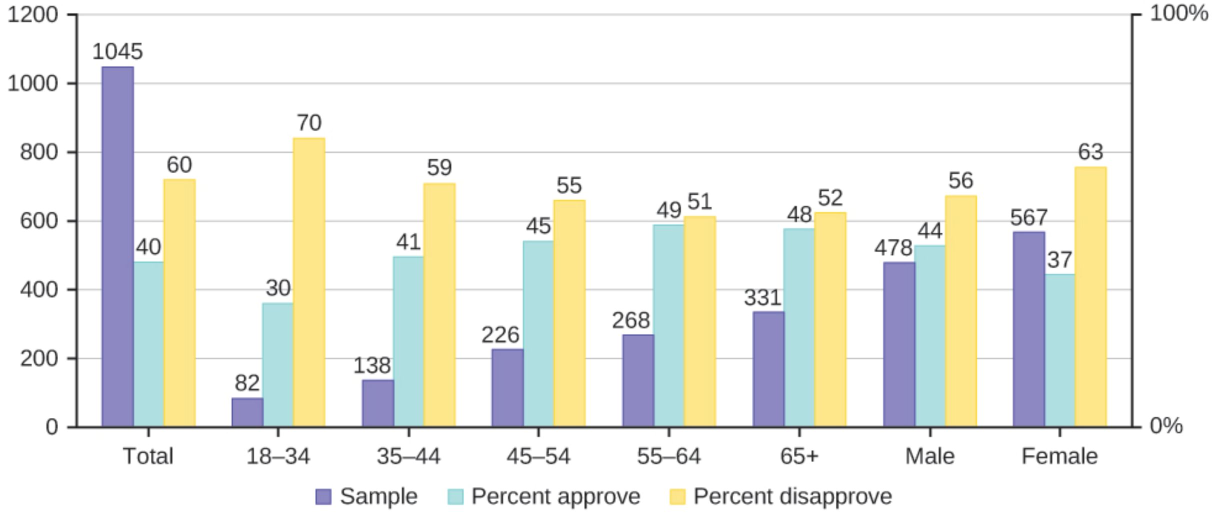


Figure 4.29. Figure description available at the end of the section.

a. Define three events in the graph.
b. Describe in words what the bar marked "40" means in this graph.
c. Describe in words the complement of the entry in (b).
d. Describe in words what the bar marked "30" means in this graph.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find P(approve|male).
i. Out of the age groups, what percent are more than 44 years old?
j. Find P(approve|age < 35).

---

39. Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130%

chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

4.2 Discrete Random Variables

1. Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. *C* = the event that Helen makes the first shot. *P*(*C*) = 0.75. *D* = the event Helen makes the second shot. *P*(*D*) = 0.75. The probability that Helen makes the second free throw, given that she made the first, is 0.85. What is the probability that Helen makes both free throws?

2. A community swim team has 150 members. Seventy-five of the members are advanced swimmers. Forty-seven of the members are intermediate swimmers. The remainder are novice swimmers. Forty of the advanced swimmers practice four times a week. Thirty of the intermediate swimmers practice four times a week. Ten of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?
b. What is the probability that the member practices four times a week?
c. What is the probability that the member is an advanced swimmer and practices four times a week?
d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?
e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

3. A school has 200 seniors, of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors who are going to college play sports. Thirty of the seniors who are going directly to work play sports. Five of the seniors who are taking a gap year play sports. What is the probability that a senior is taking a gap year?

4. A student goes to the library. Let events *B* = the student checks out a book and *D* = the student check out a DVD. Suppose that *P*(*B*) = 0.40, *P*(*D*) = 0.30 and *P*(*D*|*B*) = 0.5.

a. Find *P*(*B* AND *D*).
b. Find *P*(*B* OR *D*).

5. Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that, of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that, in the general population of women, the test for breast cancer is negative about 85% of the time. Let *B* = woman develops breast cancer, and let *N* = tests negative. Suppose one woman is selected at random.

a. What is the probability that the woman develops breast cancer? What is the probability that the woman tests negative?
b. Given that the woman has breast cancer, what is the probability that she tests negative?
c. What is the probability that the woman has breast cancer AND tests negative?
d. What is the probability that the woman has breast cancer or tests negative?
e. Are having breast cancer and testing negative independent events?
f. Are having breast cancer and testing negative mutually exclusive?

---

6. A school has 200 seniors, of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors who are going to college play sports. Thirty of the seniors who are going directly to work play sports. Five of the seniors who are taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

---

7. Refer to the information in Question 5. Let *P* = tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(P|B) = 1 - P(N|B)$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(B \text{ AND } P) = P(P|B)P(B)$?
c. What is the probability that a woman does not develop breast cancer. Find $P(B') = 1 - P(B)$?
d. What is the probability that a woman tests positive for breast cancer. Find $P(P) = 1 - P(N)$?

---

8. A student goes to the library. Let events *B* = the student checks out a book and *D* = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$, and $P(D|B) = 0.5$.

a. Find $P(B')$.
b. Find $P(D \text{ AND } B)$.
c. Find $P(B|D)$.
d. Find $P(D \text{ AND } B')$.
e. Find $P(D|B')$.

---

9. Forty-eight percent of all Californian registered voters prefer life in prison without parole over the death penalty for a person convicted of first-degree murder. Among Latino Californian registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first-degree murder. Of all Californians, 37.6% are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first-degree murder
- L = Latino Californians

Suppose that one Californian is randomly selected.

a. Find *P*(C).
b. Find *P*(L).
c. Find *P*(C|L).
d. In words, what is C|L?
e. Find *P*(L AND C).
f. In words, what is L AND C?
g. Are L and C independent events? Show why or why not.
h. Find *P*(L OR C).
i. In words, what is L OR C?
j. Are L and C mutually exclusive events? Show why or why not.

10. On February 28, 2013, a Field Poll Survey reported that 61% of Californian registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18-to-39-year-old Californian registered voters, the approval rating was 78%. Six in ten Californian registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.[2]

In this problem, let:

- C = California registered voters who support same-sex marriage
- B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- A = California registered voters who are 18-to-39 years old

a. Find *P*(C).

b. Find P(B).
c. Find P(C|A).
d. Find P(B|C).
e. In words, what is C|A?
f. In words, what is B|C?
g. Find P(C AND B).
h. In words, what is C AND B?
i. Find P(C OR B).
j. Are C and B mutually exclusive events? Show why or why not.

---

11. After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. [3]

These are the results their poll produced:

- In early 2011, 60% of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57% of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42%.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford based on the data collected in mid-2011?
e. What is the probability that a person supported Mayor Ford based on the data collected in early 2011?

12. The casino game roulette allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains 38 numbers, and each number is assigned to a color and a range.[4]

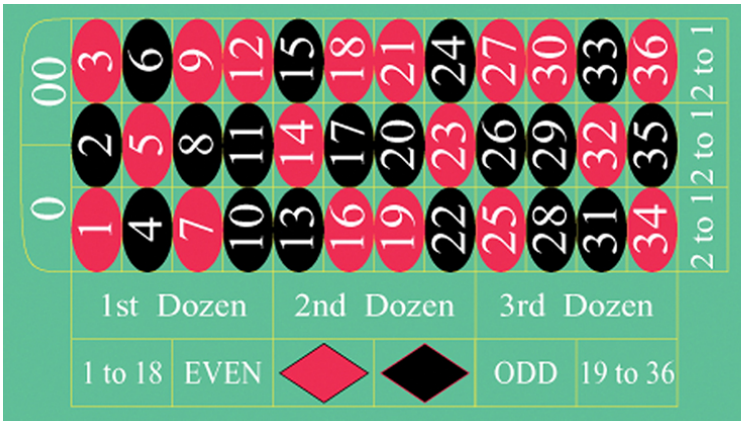


Figure 4.30. Figure description available at the end of the section.

Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table, as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

---

13. Suppose that you have eight cards. Five are green, and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

a. List the sample space.
b. Find *P*(G.
c. Find *P*(G|E).
d. Find *P*(G AND E).
e. Find *P*(G OR E).
f. Are G and E mutually exclusive? Justify your answer numerically.

14. Roll two fair dice separately. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first followed by an even number. Find *P*(A).
c. Let B be the event that the sum of the two rolls is at most seven. Find *P*(B).
d. In words, explain what "P(A|B)" represents. Find P(A|B).
e. Are A and B mutually exclusive events? Explain your answer in one-to-three complete sentences including numerical justification.
f. Are A and B independent events? Explain your answer in one-to-three complete sentences including numerical justification.

---

15. A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find P(A).

c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one-to-three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one-to-three complete sentences, including numerical justification.

NOTE: The coin toss is independent of the card picked first.

---

16. An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first followed by landing a head on the coin toss. Find *P*(A).
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one-to-three complete sentences including numerical justification.

---

17. An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side on which the coin lands.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find *P*(A).
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one-to-three complete sentences, including justification.

---

18. Consider the following scenario:

- Let *P*(C) = 0.4.
- Let *P*(*D*) = 0.5.
- Let *P*(C|*D*) = 0.6.

a. Find *P*(C AND *D*).
b. Are C and *D* mutually exclusive? Why or why not?
c. Are C and *D* independent events? Why or why not?
d. Find *P*(C OR *D*).
e. Find *P*(*D*|C).

---

19. Y and Z are independent events.

a. Rewrite the basic addition rule, P(Y OR Z) = P(Y) + P(Z) − P(Y AND Z), using the information that Y and Z are independent events.
b. Use the rewritten rule to find P(Z) if P(Y OR Z) = 0.71 and P(Y) = 0.42.

---

20. G and H are mutually exclusive events. P(G) = 0.5 P(H) = 0.3.

a. Explain why the following statement MUST be false: P(H|G) = 0.4.
b. Find P(H OR G).
c. Are G and H independent or dependent events? Explain in a complete sentence.

---

21. Approximately 281,000,000 people over the age of five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.[5]

- Let E = speaks English at home.
- Let E′ = speaks another language at home.
- Let S = speaks Spanish.

Finish each probability statement in the table below by matching the correct answer.

Probability statements	Answers
a. $P(E')$ =	i. 0.8043
b. $P(E)$ =	ii. 0.623
c. $P(S \text{ and } E')$ =	iii. 0.1957
d. $P(S \mid E')$ =	iv. 0.1219

Figure 4.31

---

22. In 1994, the US government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the US). Renate Deutsch from Germany was one of approximately 6.5 million people who entered this lottery. Let G = won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance

of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.

c. Are G and F independent or dependent events? Justify your answer numerically, and explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically, and explain why.

23. Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. Overall, 44% were returned. From the economics classes, 56% of the envelopes were returned. From the business, psychology, and history classes, 31% were returned.

- Let R = money returned.
- Let E = economics classes.
- Let O = other classes.

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.
c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically, and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

24. The following table of data obtained from the website Baseball Almanac shows hit information for four players. Suppose that one hit from the table is randomly selected.[6]

Name	Single	Double	Triple	Home run	Total hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Figure 4.32

Are “the hit being made by Hank Aaron” and “the hit being a double” independent events?

a. Yes, because P(hit by Hank Aaron|hit is a double) = P(hit by Hank Aaron).
b. No, because P(hit by Hank Aaron|hit is a double) ≠ P(hit is a double).
c. No, because P(hit is by Hank Aaron|hit is a double) ≠ P(hit by Hank Aaron).
d. Yes, because P(hit is by Hank Aaron|hit is a double) = P(hit is a double).

25. United Blood Services[7] is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Their data show that 43% of people have type O blood, 15% of people have Rh- factor, and 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

---

26. At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let *F* be the event that a course has a final exam. Let *R* be the event that a course requires a research paper.

a. Find the probability that a course has a final exam or a research project.
b. Find the probability that a course has NEITHER of these two requirements.

27. In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

a. Find the probability that a cookie contains chocolate or nuts (i.e., he can't eat it).
b. Find the probability that a cookie does not contain chocolate or nuts (i.e., he can eat it).

---

28. A college finds that 10% of students have taken a distance learning class and that 40% of students are part-time students. Of the part time students, 20% have taken a distance learning class. Let *D* = the event that a student takes a distance learning class and E = the event that a student is a part-time student.

a. Find *P*(*D* AND *E*).
b. Find *P*(*E*|*D*).
c. Find *P*(*D* OR *E*).
d. Using an appropriate test, show whether *D* and E are independent.
e. Using an appropriate test, show whether *D* and E are mutually exclusive.

---

29. A company wants to evaluate its attrition rate—in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution. Let X = the number of

years a new hire will stay with the company. Let P(*x*) = the probability that a new hire will stay with the company *x* years. Complete the figure below using the data provided.

x	P(*x*)
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

Figure 4.33

a. Find $P(x = 4)$.
b. Find $P(x \geq 5)$.
c. On average, how long would you expect a new hire to stay with the company?
d. What does the column "$P(x)$" sum to?

30. A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	P(*x*)
1	0.15
2	0.35
3	0.40
4	0.10

Figure 4.34

a. Define the random variable X.
b. What is the probability the baker will sell more than one batch? $P(x > 1)$
c. What is the probability the baker will sell exactly one batch? $P(x = 1)$
d. On average, how many batches should the baker make?

31. Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

a. Define the random variable X.

b. Construct a probability distribution table for the data.
c. We know that for a probability distribution function must have two characteristics to be discrete. One is that the sum of the probabilities is one. What is the other characteristic?

---

32. Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

a. Define the random variable X.
b. What values does *x* take on?
c. Construct a PDF table.
d. Find the probability that Javier volunteers for less than three events each month. $P(x < 3)$
e. Find the probability that Javier volunteers for at least one event each month. $P(x > 0)$

33. Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (BS) degree is given in below.

x	P(*x*)
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

Figure 4.35

a. In words, define the random variable X.
b. What does it mean that the values 0, 1, and 2 are not included for *x* in the PDF?

---

34. Suppose you play a game of chance in which a computer randomly chooses five numbers from 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 with replacement. You pay $2 to play and could profit $100,000 if you match all five numbers in order (getting your $2 back plus $100,000). Over the long term, what is your expected profit of playing the game?

- To do this problem, set up an expected value table for the amount of money you can profit. Let X = the amount of money you profit. The values of *x* are not 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Since you are interested in your profit (or loss), the values of *x* are 100,000 dollars and −2 dollars.
- To win, you must get all five numbers correct, in order. The probability of choosing one correct number

is $\frac{1}{10}$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is
$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = (1)\left(10^{-5}\right) = 0.00001$

- Therefore, the probability of winning is 0.00001 and the probability of losing is 1 – 0.00001 = 0.99999.
- The expected value table is as follows: (Add the last column. –1.99998 + 1 = –0.99998.)

	x	P(x)	x*P(x)
Loss	-2	0.99999	(–2)(0.99999) = –1.99998
Profit	100,000	0.00001	(100000)(0.00001) = 1

Figure 4.36

Since –0.99998 is about –1, you would, on average, expect to lose approximately $1 for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected loss per game after playing this game over and over.

35. You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay $1 to play. If you guess the right suit every time, you get your money back and $256. What is your expected profit from playing the game over the long term?

---

36. Suppose you play a game in which you toss a biased coin once, with the probabilities P(heads) = $\frac{2}{3}$ and P(tails) = $\frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

a. Define a random variable X.
b. Complete the following expected value table:

	x		
Win	10	$\frac{1}{3}$	
Lose			$\frac{-12}{3}$

Figure 4.37

c. What is the expected value, μ? Do you come out ahead?

---

37. Suppose you play a game in which you spin a spinner once, with the probabilities P(red) = $\frac{2}{5}$, P(blue) = $\frac{2}{5}$, and P(green) = $\frac{1}{5}$. If you land on red, you pay $10. If you land on blue, you don't pay or win anything. If you land on green, you win $10. Complete the following expected value table:

	x	P(x)	
Red			$\frac{-20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Figure 4.38

---

38. Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like the one in Question 4, and calculate the mean (μ) and standard deviation (σ) of X.

Tossing one fair, six-sided die twice has the same sample space as tossing two fair, six-sided dice. The sample space has 36 outcomes:

Sample space outcomes					
(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Figure 4.39

a. Use the sample space to complete the following table:

x	P(x)	x*P(x)	$(x-\mu)^2$*P(x)
0			
1			
2			

Figure 4.40

b. Add the values in the third column to find the expected value. Use this value to complete the fourth column.

c. Add the values in the fourth column and take the square root of the sum.

---

39. On May 11, 2013, at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earth-

quake will occur in Iran during this period. If you win the bet, you win $50. If you lose the bet, you pay $20. Let X = the amount of profit from a bet.[8]

- P(win) = P(one moderate earthquake will occur) = 21.42%
- P(loss) = P(one moderate earthquake will *not* occur) = 100% − 21.42%

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of X? Construct a table similar to those in Questions 4 and 5 to help you answer these questions.

---

40. Complete the expected value table.

x	P(x)	x*P(x)
0	0.2	
1	0.2	
2	0.4	
3	0.2	

Figure 4.41

---

41. Find the expected value from the expected value table. Find the standard deviation.

x	P(x)	x*P(x)
2	0.1	2(0.1) = 0.2
4	0.3	4(0.3) = 1.2
6	0.4	6(0.4) = 2.4
8	0.2	8(0.2) = 1.6

Figure 4.42

---

42. Identify the mistake in the probability distribution table.

x	P(x)	x*P(x)
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

Figure 4.43

43. Identify the mistake in the probability distribution table.

x	P(x)	x*P(x)
1	0.15	0.15
2	0.25	0.40
3	0.25	0.65
4	0.20	0.85
5	0.15	1

Figure 4.44

44. A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution:

x	P(x)	x*P(x)
1	0.35	
2	0.20	
3	0.15	
4		
5	0.10	
6	0.50	

Figure 4.45

a. Define the random variable X.
b. Define *P*(*x*), or the probability of *x*.
c. Find the probability that a physics major will do post-graduate research for four years. *P*(*x* = 4)
d. Find the probability that a physics major will do post-graduate research for at most three years. *P*(*x* ≤ 3)
e. On average, how many years would you expect a physics major to spend doing post-graduate research?

45. A ballet instructor is interested in knowing what percent of each year's class will continue on to the next so that she can plan what classes to offer. Over the years, she has established a probability distribution.

- Let X = the number of years a student will study ballet with the teacher.
- Let *P*(*x*) = the probability that a student will study ballet *x* years.

Complete the figure below using the data provided.

x	*P*(*x*)	*x*P*(*x*)
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

Figure 4.46

a. In words, define the random variable X.
b. Find $P(x = 4)$.
c. Find $P(x < 4)$.
d. On average, how many years would you expect a child to study ballet with this teacher?
e. What does the column "$P(x)$" sum to and why?
f. What does the column "$x*P(x)$" sum to and why?

46. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards.

a. What is the expected value of playing the game?
b. Should you play the game?

47. A theater group holds a fundraiser. It sells 100 raffle tickets for $5 apiece. The prize is two passes to a Broadway show, worth a total of $150. Suppose you purchase four tickets.

a. What are you interested in here?
b. In words, define the random variable X.
c. List the values that X may take on.
d. Construct a PDF.
e. If this fundraiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?

48. A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on heads, you win $6.
- If the card is a face card, and the coin lands on tails, you win $2.
- If the card is not a face card, you lose $2, no matter what the coin shows.

a. Find the expected value for this game (expected net gain or loss).
b. Explain what your calculations indicate about your long-term average profits and losses on this game.
c. Should you play this game to win money?

---

49. You buy a lottery ticket to a lottery that costs $10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery, there are one $500 prize, two $100 prizes, and four $25 prizes. Find your expected gain or loss.

- Start by writing the probability distribution. X is net gain or loss = prize (if any) less $10 cost of ticket.
- Expected value = (490)(1,100) + (90)(2,100) + (15)(4,100) + (–10)(93,100) = –$2. There is an expected loss of $2 per ticket, on average.

Complete the PDF and answer the questions.

x	P(x)	xP(x)
0	0.3	
1	0.2	
2		
3	0.4	

Figure 4.47

a. Find the probability that *x* = 2.
b. Find the expected value.

---

50. Suppose that you are given a die to roll. If you roll a six, you win $10. If you roll a four or five, you win $5. If you roll a one, two, or three, you pay $6.

a. What are you ultimately interested in here (the value of the roll or the money you win)?
b. In words, define the random variable, X.
c. List the values that X may take on.
d. Construct a PDF.
e. Over the long run of playing this game, what are your expected average winnings per game?
f. Based on numerical values, should you take the deal? Explain your decision in complete sentences.

---

51. A venture capitalist willing to invest $1,000,000 has three investments from which to choose. The first investment, a software company, has a 10% chance of returning $5,000,000 profit, a 30% chance of returning $1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning $3,000,000 profit, a 40% chance of returning $1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning $6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

a. Construct a PDF for each investment.
b. Find the expected value for each investment.
c. Which is the safest investment? Why do you think so?
d. Which is the riskiest investment? Why do you think so?
e. Which investment has the highest expected return, on average?

52. Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X = the number of children married people have.

x	P(x)	x*P(x)
0	0.10	
1	0.20	
2	0.30	
3		
4	0.10	
5	0.05	
6 (or more)	0.05	

Figure 4.48

a. Find the probability that a married adult has three children.
b. In words, what does the expected value in this example represent?
c. Find the expected value.
d. Is it more likely that a married adult will have two-to-three children or four-to-six children? How do you know?

---

53. Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (BS) degree is given below.

x	P(x)
3	0.05
4	0.40

x	P(*x*)
5	0.30
6	0.15
7	0.10

Figure 4.49

On average, how many years do you expect it to take for an individual to earn a BS?

---

54. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	P(*x*)
0	0.30
1	0.50
2	0.24
3	
4	0.07
5	0.04

Figure 4.50

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.
e. Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

x	P(*x*)
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

Figure 4.51

At which store is the expected number of DVDs rented per customer higher?

f. If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
g. If Video to Go expects 300 customers next week, and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
h. Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

---

55. A friend offers you the following deal." For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.
b. No, I expect to come out behind in money.
c. It doesn't matter. I expect to break even.

---

56. Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.[9]

a. What is the average class size assuming each class is filled to capacity?
b. Space is available for 980 students. Suppose that each class is filled to capacity, and select a statistics student at random. Let the random variable X equal the size of the student's class. Define the PDF for X.
c. Find the mean of X.
d. Find the standard deviation of X.

---

57. In a lottery, there are 250 prizes of $5, 50 prizes of $25, and ten prizes of $100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

Let X = the amount of money to be won on a ticket. The following table shows the PDF for X.

x	P(*x*)
0	0.969

x	P(x)
5	$\frac{250}{10{,}000} = 0.025$
25	$\frac{50}{10{,}000} = 0.005$
100	$\frac{10}{10{,}000} = 0.001$

Figure 4.52

Calculate the expected value of X.

4.3 The Binomial Distribution

1. The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a "success" be in this case?

---

2. A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

---

3. A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

---

4. The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%).[10] Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Find the probability that at most eight people develop pancreatic cancer.
d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

---

5. During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league, scoring with 61.3% of his shots.[11] Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Find the probability that DeAndre scored with 60 of these shots.
d. Find the probability that DeAndre scored with more than 50 of these shots.

6. A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be captain twice). You want to see if the captains all play the same position. State whether this is binomial or not, and state why.

---

7. The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the US. Of those students, 71.3% replied that, yes, they believe that same-sex couples should have the right to legal marital status.[12] Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

a. In words, define the random variable X.
b. X ~ ___(__,__)
c. What values does the random variable X take on?
d. Construct the probability distribution function (PDF).
e. On average (μ), how many would you expect to answer yes?
f. What is the standard deviation (σ)?
g. What is the probability that at most five of the freshmen reply “yes”?
h. What is the probability that at least two of the freshmen reply “yes”?

---

8. According to a recent article, the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

---

9. Recently, a nurse commented that, when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

a. Define the random variable, and list its possible values.
b. State the distribution of X.
c. Find the probability that at least four of the 25 patients actually have the flu.
d. On average, for every 25 patients calling in, how many do you expect to have the flu?

10. A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. How many of the 12 students do we expect to attend the festivities?
e. Find the probability that at most four students will attend.
f. Find the probability that more than two students will attend.

---

11. The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date).[13] An upcoming monthly schedule contains 12 games.

a. The expected number of wins for that upcoming month is:

a. 1.67
b. 12
c. $\frac{382}{1034}$
d. 4.43

b. Let X = the number of games won in that upcoming month. What is the probability that the San Jose Sharks win six games in that upcoming month?

a. 0.1476
b. 0.2336
c. 0.7664
d. 0.8903

c. What is the probability that the San Jose Sharks win at least five games in that upcoming month?

a. 0.3694
b. 0.5266
c. 0.4734
d. 0.2305

---

12. A student takes a ten-question true-false quiz but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70%.

13. A student takes a 32-question multiple-choice exam but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses more than 75% of the questions correctly.

---

14. Six different colored dice are rolled. Of interest is the number of dice that show a one.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ __(__,__)
d. On average, how many dice would you expect to show a one?
e. Find the probability that all six dice show a one.
f. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

---

15. More than 96% of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. On average, how many schools would you expect to offer such courses?
e. Find the probability that at most ten offer such courses.
f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

---

16. Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. How many are expected to attend their graduation?
e. Find the probability that 17 or 18 attend.
f. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

17. At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do not use the foil as their main weapon.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. How many are expected to not to use the foil as their main weapon?
e. Find the probability that six do not use the foil as their main weapon.
f. Based on numerical values, would you be surprised if all 25 did not use foil as their main weapon? Justify your answer numerically.

---

18. Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based on numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

---

19. The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. We are

interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. How many audits are expected in a 20-year period?
e. Find the probability that a person is not audited at all.
f. Find the probability that a person is audited more than twice.

20. It has been estimated that only about 30% of California residents have adequate earthquake supplies.[14] Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. What is the probability that at least eight have adequate earthquake supplies?
e. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
f. How many residents do you expect will have adequate earthquake supplies?

---

21. There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used, along with a board with those six objects on it. We will play with bets being $1. The player places a bet on a number or object. The "house" rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back their $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back their $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back their $1 bet, plus $3 profit. Let X = number of matches and Y = profit per game.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ___(__,__)
d. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.

e. Calculate the average expected matches over the long run of playing this game for the player.
f. Calculate the average expected earnings over the long run of playing this game for the player.
g. Determine who has the advantage, the player or the house.

---

22. According to the World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X = the number of people who have access to electricity.

a. What is the probability distribution for X?
b. Using the formulas, calculate the mean and standard deviation of X.
c. Find the probability that 15 people in the sample have access to electricity.
d. Find the probability that at most ten people in the sample have access to electricity.
e. Find the probability that more than 25 people in the sample have access to electricity.

---

23. The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let X = the number of people who are literate.[15]

a. Sketch a graph of the probability distribution of X.
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Find the probability that more than five people in the sample are literate. Is it is more likely that three people or four people are literate?

---

24. Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times. Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, …, 20. The probability of a success is p = 0.55. The probability of a failure is q = 0.45. The number of trials is n = 20. The probability question can be stated mathematically as P(x = 15).

---

25. A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, …, 15. Since the coin is fair, p = 0.5 and q = 0.5. The number of trials is n = 15. State the probability question mathematically.

4.4 Continuous Random Variables

1. What does the shaded area represent? $P(< x <)$

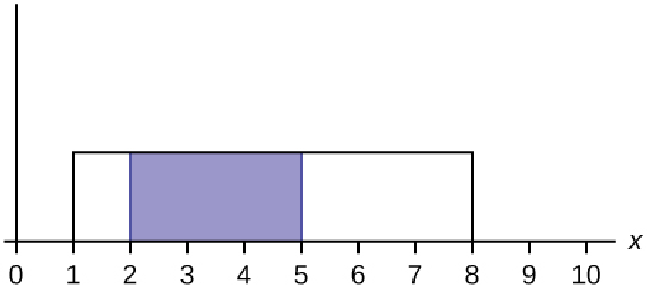


Figure 4.53. *Figure description available at the end of the section*.

---

2. What does the shaded area represent? $P(< x <)$

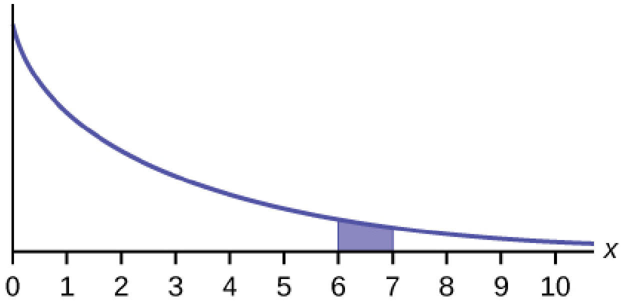


Figure 4.54. *Figure description available at the end of the section*.

---

3. For a continuous probability distribution, $0 \le x \le 15$. What is $P(x > 15)$?

4. What is the area under $f(x)$ if the function is a continuous probability density function?

5. For a continuous probability distribution, $0 \le x \le 10$. What is $P(x = 7)$?

6. A continuous probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?

7. $f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \le x \le 5$. What is $P(x < 0)$?

8. $f(x)$ for a continuous probability function is equal to $\frac{1}{12}$, and the function is restricted to $0 \le x \le 12$. What is $P\ (0 < x < 12)$?

9. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an RN (registered nurse) degree. You ask nurses if they have an RN degree. The nurses answer "yes" or "no." You then calculate the percentage of nurses with an RN degree. You give that percentage to your supervisor.

a. What part of the experiment will yield discrete data?
b. What part of the experiment will yield continuous data?
c. When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?

4.5 The Normal Distribution

1. The mean height of 15- to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm.[16] Male heights are known to follow a normal distribution. Let X = the height of a 15- to 18-year-old male from Chile in 2009 to 2010. Then, X ~ N(170, 6.28).

a. Suppose a 15- to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when x = 168 cm is z = . This z-score tells you that x = 168 is standard deviations to the (right or left) of the mean, which is .
b. Suppose that the height of a 15- to 18-year-old male from Chile from 2009 to 2010 has a z-score of z = 1.27. What is the male's height? The z-score (z = 1.27) tells you that the male's height is standard deviations to the (right or left) of the mean.

2. Use the information in Question 1 to answer the following questions.

a. Suppose a 15- to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when x = 176 cm is z = . This z-score tells you that x = 176 cm is standard deviations to the (right or left) of the mean, which is .
b. Suppose that the height of a 15- to 18-year-old male from Chile from 2009 to 2010 has a z-score of z = −2. What is the male's height? The z-score (z = −2) tells you that the male's height is standard deviations to the (right or left) of the mean.

3. In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean μ = 496 and a standard deviation σ = 114.[17] Let X = a SAT exam verbal section score in 2012. Then, X ~ N(496, 114). Find the z-scores for x_1 = 325 and x_2 = 366.21. Interpret each z-score. What can you say about x_1 = 325 and x_2 = 366.21 in comparison to their respective means and standard deviations?

4. What is the z-score of x when x = 1 and X ~ N(12, 3)?

5. Some doctors believe that a person can lose five pounds, on the average, in a month by reducing their fat intake and by exercising consistently.[18] Suppose weight loss has a normal distribution. Let X = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of two pounds. X ~ N(5, 2). Fill in the blanks.

a. Suppose a person lost ten pounds in a month. The z-score when x = 10 pounds is z = 2.5 (verify). This z-score tells you that x = 10 is [?] standard deviations to the [right/left] of the mean, which is [?].
b. Suppose a person gained three pounds (a negative weight loss). Then z = [?]. This z-score tells you that x = −3 is [?] standard deviations to the [right/left] of the mean.
c. Suppose the random variables X and Y have the following normal distributions: X ~ N(5, 6) and Y ~ N(2, 1). If x = 17, then z = 2; if y = 4, what is z?

6. Jerome averages 16 points a game with a standard deviation of four points. X ~ N(16, 4). Suppose Jerome scores ten points in a game. The z–score when x = 10 is –1.5. This score tells you that x = 10 is [?] standard deviations to the [right/left] of the mean, which is [?].

7. From 1984 to 1985, the mean height of 15- to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm.[19] Let Y = the height of 15- to 18-year-old males from 1984 to 1985. Then, Y ~ N(172.36, 6.34).

The mean height of 15- to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15- to 18-year-old male from Chile in 2009 to 2010. Then, X ~ N(170, 6.28).

Find the z-scores for x = 160.58 cm and y = 162.85 cm. Interpret each z-score. What can you say about x = 160.58 cm and y = 162.85 cm in comparison to their respective means and standard deviations?

8. In 2012, 1,664,479 students took the SAT exam.[20] The distribution of scores in the verbal section of the SAT had a mean μ = 496 and a standard deviation σ = 114. Let X = a SAT exam verbal section score in 2012. Then, X ~ N(496, 114). Find the z-scores for x_1 = 325 and x_2 = 366.21. Interpret each z-score. What can you say about x_1 = 325 and x_2 = 366.21 in comparison to their respective means and standard deviations?

9. Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

10. The scores on a college entrance exam have an approximate normal distribution with mean μ = 52 points and a standard deviation σ = 11 points.

a. About 68% of the y values lie between what two values? What are the respective z-scores?
b. About 95% of the y values lie between what two values? What are the respective z-scores?
c. About 99.7% of the y values lie between what two values? What are the respective z-scores?

11. A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words.

12. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

13. X ~ N(1, 2). σ =

14. A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words.

15. X ~ N(−4, 1). What is the median?

16. X ~ N(3, 5). σ =

17. X ~ N(−2, 1). μ =

18. What does a *z*-score measure?

19. What does standardizing a normal distribution do to the mean?

20. Is X ~ N(0, 1) a standardized normal distribution? Why or why not?

21. What is the *z*-score of x = 12 if it is two standard deviations to the right of the mean?

22. What is the *z*-score of x = 9 if it is 1.5 standard deviations to the left of the mean?

23. What is the z-score of $x = -2$ if it is 2.78 standard deviations to the right of the mean?

24. What is the z-score of $x = 7$ if it is 0.133 standard deviations to the left of the mean?

25. Suppose X ~ N(2, 6). What value of x has a z-score of three?

26. Suppose X ~ N(8, 1). What value of x has a z-score of −2.25?

27. Suppose X ~ N(9, 5). What value of x has a z-score of −0.5?

28. Suppose X ~ N(2, 3). What value of x has a z-score of −0.67?

29. Suppose X ~ N(4, 2). What value of x is 1.5 standard deviations to the left of the mean?

30. Suppose X ~ N(4, 2). What value of x is two standard deviations to the right of the mean?

31. Suppose X ~ N(8, 9). What value of x is 0.67 standard deviations to the left of the mean?

32. Suppose X ~ N(−1, 2). What is the z-score of $x = 2$?

33. Suppose X ~ N(12, 6). What is the z-score of $x = 2$?

34. Suppose X ~ N(9, 3). What is the z-score of $x = 9$?

35. Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of x = 5.5?

36. In a normal distribution, x = 5 and z = −1.25. This tells you that x = 5 is standard deviations to the (right or left) of the mean.

37. In a normal distribution, x = 3 and z = 0.67. This tells you that x = 3 is standard deviations to the (right or left) of the mean.

38. In a normal distribution, x = −2 and z = 6. This tells you that x = −2 is standard deviations to the (right or left) of the mean.

39. In a normal distribution, x = −5 and z = −3.14. This tells you that x = −5 is standard deviations to the (right or left) of the mean.

40. In a normal distribution, x = 6 and z = −1.7. This tells you that x = 6 is standard deviations to the (right or left) of the mean.

41. About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

42. About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

43. About what percent of x values lie between the second and third standard deviations (both sides)?

44. Suppose X ~ N(15, 3). Between what *x* values does 68.27% of the data lie? The range of *x* values is centered at the mean of the distribution (i.e., 15).

45. Suppose X ~ N(−3, 1). Between what *x* values does 95.45% of the data lie? The range of *x* values is centered at the mean of the distribution (i.e., −3).

---

46. Suppose X ~ N(−3, 1). Between what *x* values does 34.14% of the data lie?

---

47. About what percent of *x* values lie between the mean and three standard deviations?

---

48. About what percent of *x* values lie between the mean and one standard deviation?

---

49. About what percent of *x* values lie between the first and second standard deviations from the mean (both sides)?

---

50. About what percent of *x* values lie between the first and third standard deviations (both sides)?

---

51. The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

a. Define the random variable X in words.
b. X ~ ___(__,__)

---

52. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

a. What is the median recovery time?
b. What is the *z*-score for a patient who takes ten days to recover?

53. The length of time to find it takes to find a parking space at 9 AM follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution.
III. The data cannot follow the normal distribution.

a. I only
b. II only
c. III only
d. I, II, and III

---

54. The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean $\mu = 79$ inches and standard deviation $\sigma = 3.89$ inches.[21] For each of the following heights, calculate the z-score and interpret it using complete sentences.

a. 77 inches
b. 85 inches
c. If an NBA player reported his height had a z-score of 3.5, would you believe him? Explain your answer.

---

55. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.[22]

a. Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.
b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

---

56. Kyle's doctor told him that the z-score for his systolic blood pressure is 1.75. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard

deviation σ = 14. If X = a systolic blood pressure score, then X ~ N (125, 14). Which of the following is the best interpretation of Kyle's standardized score?

a. Which answer(s) is/are correct?
   i. Kyle's systolic blood pressure is 175.
   ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
   iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
   iv. Kyle's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
b. Calculate Kyle's blood pressure.

---

57. In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean μ = 520 and standard deviation σ = 115.[23]

a. Calculate the *z*-score for an SAT score of 720. Interpret it using a complete sentence.
b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
c. For 2012, the SAT math test had a mean of 514 and standard deviation of 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3.[24] If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

4.6 The Normal Approximation to the Binomial

1. Suppose in a local kindergarten through 12th grade (K-12) school district, 53% of the population favor a charter school for Grades K through 5. A simple random sample of 300 is surveyed.

a. Find the probability that at least 150 favor a charter school.
b. Find the probability that at most 160 favor a charter school.
c. Find the probability that more than 155 favor a charter school.
d. Find the probability that fewer than 147 favor a charter school.
e. Find the probability that exactly 175 favor a charter school.

---

2. In a city, 46% of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

Figure Descriptions

Figure 4.25: Horizontal bar chart detailing emotional health score index with occupation on the y axis and values 77-85 on the x axis. In ascending order: service, transportation, manufacturing or production, sales, clerical or office, installation or repair, construction or mining, manager/executive/official, business owner, nurse, professional, farming/fishing/forestry, K-12 teacher, physician

Figure 4.28: Pie chart: four red, two blue, one yellow, one green

Figure 4.29: Total sample: 1045, total percent approve: 40, total percent disapprove: 60. 18-34 sample: 82, 18-34 percent approve: 30, 18-38 percent disapprove: 70. 35-44 sample: 138, 35-44 percent approve: 41, 35-44 percent disapprove: 59. 45-54 sample: 226, 45-54 percent approve: 45, 45-54 percent disapprove: 55. 55-64 sample: 268, 55-64 percent approve: 49, 55-64 percent disapprove: 51. 65+ sample: 331, 65+ percent approve: 48, 65+ percent disapprove: 52. Male sample: 478, Male percent approve: 44, Male percent disapprove: 56. Female sample: 567, Female percent approve: 37, Female percent disapprove: 63.

Figure 4.30: Roulette table with numbers one through 36 in alternating red and black colors (starting with red).

Figure 4.53: This graph shows a uniform distribution. The horizontal axis ranges from zero to 10. The distribution is modeled by a rectangle extending from x = one to x = eight. A region from x = two to x = five is shaded inside the rectangle.

Figure 4.54: This graph shows an exponential distribution. The graph slopes downward. It begins at a point on the y-axis and approaches the x-axis at the right edge of the graph. The region under the graph from x = six to x = seven is shaded.

References

Figures

Figure 4.25: Figure 3.15.18 from LibreTexts Business Statistics (2021) (CC BY 4.0). Retrieved from https://stats.libretexts.org/Bookshelves/Applied_Statistics/Business_Statistics_(OpenStax)/03%3A_Probability_Topics/3.15%3A_Homework

Figure 4.28: Figure 3.10 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/3-practice

Figure 4.29: Figure 3.11 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/3-homework

Figure 4.30: Figure 3.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics/pages/3-homework

Figure 4.53: Figure 5.26 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-business-statistics/pages/5-practice

Figure 4.54: Figure 5.27 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-business-statistics/pages/5-practice

Text

Class Catalogue at the Florida State University. Available online at https://apps.oti.fsu.edu/Registrar-CourseLookup/SearchFormLegacy (accessed May 15, 2013).

"World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

"Access to electricity (% of population)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

"Distance Education," Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

"NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/season-type/2 (accessed May 15, 2013).

Newport, Frank. "Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income," *Gallup Economy*, 2013. Available online at http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, and Serge Tran. "The American Freshman: National Norms Fall 2011." Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf (accessed May 15, 2013).

"The World FactBook," Central Intelligence Agency. Available online at https://www.cia.gov/library/publications/the-world-factbook/geos/af.html (accessed May 15, 2013).

"What are the key statistics about pancreatic cancer?" American Cancer Society, 2013. Available online at http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics (accessed May 15, 2013).

"Blood Pressure of Males and Females," StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).

"The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores," London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

"2012 College-Bound Seniors Total Group Profile Report," CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).

"Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009," National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

"List of stadiums by capacity," Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Notes

1. Data from *San Jose Mercury News*.
2. DiCamillo, Mark, and Mervin Field. "The File Poll," Field Research Corporation. Available online at https://web.archive.org/web/20130512064934/http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf (accessed May 12, 2013).
3. Rider, David. "Ford support plummeting, poll suggests," *The Star*, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).
4. "Roulette." Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).
5. Shin, Hyon B., and Robert A. Kominski. "Language Use in the United States: 2007." United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf (accessed May 2, 2013).
6. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).
7. "Human Blood Types," United Blood Services, 2011. Available online at https://web.archive.org/web/20130807103902/http://www.unitedbloodservices.org/learnMore.aspx (accessed August 22, 2013).
8. "World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. http://www.worldearthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).
9. Class Catalogue at the Florida State University. Available online at https://apps.oti.fsu.edu/RegistrarCourseLookup/SearchFormLegacy (accessed May 15, 2013).
10. "What are the key statistics about pancreatic cancer?" American Cancer Society, 2013. Available online at http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics (accessed May 15, 2013).
11. "NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).
12. Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, and Serge Tran. "The American Freshman: National Norms Fall 2011." Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAm-

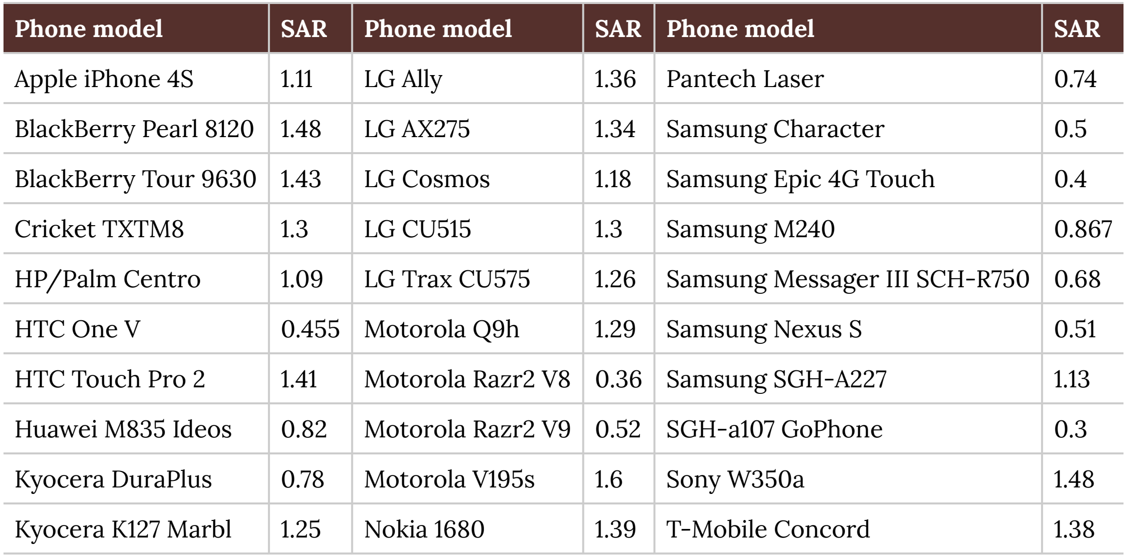
ericanFreshman2011.pdf (accessed May 15, 2013).

13. "San Jose Sharks History," Hockey Reference. Available online at https://www.hockey-reference.com/teams/SJS/history.html (accessed January 26, 2021).
14. "World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. http://www.worldearthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).
15. UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [sic] and writing skills," UNICEF Television. Video available online at http://www.unicefusa.org/assets/video/afghan-femaleliteracy-centers.html (accessed May 15, 2013).
16. Data from *The World Almanac and Book of Facts*.
17. "2012 College-Bound Seniors Total Group Profile Report," CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).
18. McDougall, John A. *The McDougall Program for Maximum Weight Loss*. Plume, 1995.
19. Data from *The World Almanac and Book of Facts*.
20. "2012 College-Bound Seniors Total Group Profile Report," CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).
21. Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).
22. "Blood Pressure of Males and Females," StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).
23. "2012 College-Bound Seniors Total Group Profile Report," CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).
24. "Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009," National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Chapter 5 Extra Practice

5.1 Point Estimation and Sampling Distributions

1. The specific absorption rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. The figure below shows the highest SAR level for a random selection of cell phone models as measured by the FCC.[1] Find a point estimate of the true (population) mean of the specific absorption rates (SARs) for cell phones.



Phone model	SAR	Phone model	SAR	Phone model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Figure 5.15

2. A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation. Find a point estimate of the true (population) proportion of students in the school district who are against the new legislation.

5.2 The Sampling Distribution of the Sample Mean (CLT)

1. The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a mean of two hours and a standard deviation of 0.5 hours. A sample of size n = 50 is drawn randomly from the population. Find the probability that the sample mean is between 1.8 hours and 2.3 hours.

- Let X = the time, in hours, it takes to play one soccer match.
- The probability question asks you to find a probability for the sample mean time, in hours, it takes to play one soccer match.
- Let $\overline{X}$ = the mean time, in hours, it takes to play one soccer match.

a. If μ_X = ______, σ_X = ______, and n = ______, then X ~ N(______, ______) by the central limit theorem for means.
b. Find $P(1.8 < \overline{x} < 2.3)$. Draw a graph.

---

2. The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours.[2] A sample size of n = 60 is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

---

3. In a recent study reported Oct. 29, 2012, on the Flurry Blog, the mean age of tablet users is 34 years.[3] Suppose the standard deviation is 15 years. Take a sample of size n = 100.

a. What are the mean and standard deviation for the sample mean ages of tablet users?
b. What does the distribution look like?
c. Find the probability that the sample mean age is more than 34 years (the reported mean age of tablet users in this particular study).
d. Find the 95th percentile for the sample mean age (to one decimal place).

---

4. An article on Flurry Blog identified a gaming marketing gap for men between the ages of 30 and 40.[4] You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

5. Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured, and the statistics are n = 34 and $\overline{x}$ = 16.01 ounces. If the cans are filled so that μ = 16.00 ounces (as labeled) and σ = 0.143 ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

---

6. Yoonie is a personnel manager in a large corporation. Each month, she must review 16 of the employees. From past experience, she has found that the reviews each take her approximately four hours to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\overline{X}$ be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

a. What is the mean, standard deviation, and sample size?
b. Complete the distributions for X and $\overline{X}$.
c. Find the probability that one review will take Yoonie from 3.5 to 4.25 hours. Sketch a graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.
d. Find the probability that the mean of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch a graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.
e. What causes the probabilities in C and D to be different?
f. Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.

---

7. Previously, De Anza statistics students estimated that the amount of change carried by daytime statistics students is exponentially distributed with a mean of 88 cents. Suppose that we randomly pick 25 daytime statistics students.

a. In words, X = ______
b. X ~ ______(______, ______)
c. In words, $\overline{X}$ = ______
d. $\overline{X}$ ~ ______ (______, ______)
e. Find the probability that an individual had between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
f. Find the probability that the average of the 25 students was between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
g. Explain why there is a difference in part (e) and (f).

---

8. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. If $\overline{X}$ = average distance in feet for 49 fly balls, then $\overline{X}$ ~ _____(_____,_____)
b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for $\overline{X}$. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of the average of 49 fly balls.

---

9. According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

a. In words, X = _____
b. In words, $\overline{X}$ = _____
c. $\overline{X}$ ~ _____(_____, _____)
d. Would you be surprised if the 36 taxpayers finished their form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
e. Would you be surprised if one taxpayer finished their form 1040 in more than 12 hours? In a complete sentence, explain why.

---

10. Suppose that a category of world-class runners are known to run a marathon (26.2 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let $\overline{X}$ the average of the 49 races.

a. $\overline{X}$ ~ _____(_____, _____)
b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
c. Find the 80th percentile for the average of these 49 marathons.
d. Find the median of the average running times.

---

11. The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

a. In words, X = _____
b. X ~ _____
c. In words, $\overline{X}$ = _____
d. $\overline{X}$ ~ _____(_____, _____)
e. Find the first quartile for the average song length, $\overline{X}$.
f. The IQR (interquartile range) for the average song length, $\overline{X}$, is from _____ – _____.

12. In 1940, the average size of a US farm was 174 acres.[5] Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

a. In words, X = ______
b. In words, $\overline{X}$ = ______
c. $\overline{X}$ ~ ______(______, ______)
d. The IQR for $\overline{X}$ is from ______ acres to ______ acres.

---

13. Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, the mean of $\overline{X}$ is approximately equal to the mean of X.
b. When the sample size is large, $\overline{X}$ is approximately normally distributed.
c. When the sample size is large, the standard deviation of $\overline{X}$ is approximately the same as the standard deviation of X.

---

14. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten.[6] Suppose that 16 individuals are randomly chosen. Let $\overline{X}$ = average percent of fat calories.

a. $\overline{X}$ ~ ______(______, ______)
b. For the group of 16, find the probability that the average percent of fat calories consumed is more than five. Graph the situation and shade in the area to be determined.
c. Find the first quartile for the average percent of fat calories.

---

15. The distribution of income in some Third World countries is considered wedge-shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge-shaped distribution. Let the average salary be $2,000 per year with a standard deviation of $8,000. We randomly survey 1,000 residents of that country.

a. In words, X = ______
b. In words, $\overline{X}$ = ______
c. $\overline{X}$ ~ ______(______, ______)
d. How is it possible for the standard deviation to be greater than the average?
e. Why is it more likely that the average of the 1,000 residents will be from $2,000 to $2,100 than from $2,100 to $2,200?

16. Which of the following is NOT TRUE about the distribution for averages?

a. The mean, median, and mode are equal.
b. The area under the curve is one.
c. The curve never touches the x-axis.
d. The curve is skewed to the right.

---

17. The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 cents and a standard deviation of 10 cents. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations. The distribution to use for the average cost of gasoline for the 16 gas stations is:

a. $\overline{X}$ ~ N(4.59, 0.10)
b. $\overline{X}$ ~ N(4.59, $\frac{0.10}{\sqrt{16}}$)
c. $\overline{X}$ ~ N(4.59, $\frac{16}{0.10}$)
d. $\overline{X}$ ~ N(4.59, $\frac{\sqrt{16}}{0.10}$)

5.3 Introduction to Confidence Intervals

1. The specific absorption rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. The figure below shows the highest SAR level for a random selection of cell phone models as measured by the FCC.[7] Find a 98% confidence interval for the true (population) mean of the specific absorption rates (SARs) for cell phones. Assume that the population standard deviation is σ = 0.337.

Phone model	SAR	Phone model	SAR	Phone model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51

Phone model	SAR	Phone model	SAR	Phone model	SAR
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Figure 5.16

2. The figure below shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States.[8] As previously, assume that the population standard deviation is σ = 0.337.

Phone model	SAR	Phone model	SAR
BlackBerry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.07	ZTE C79	0.869

Figure 5.17

3. The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

a. Identify the following:

- $\overline{x}$
- σ
- n

b. In words, define the random variables X and $\overline{X}$.
c. Which distribution should you use for this problem?
d. Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the margin of error.
e. What will happen to the confidence interval obtained if 500 newborn elephants are weighed instead of 50? Why?

---

4. The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.[9]

a. Identify the following:
   - $\overline{x}$
   - σ
   - n

b. In words, define the random variables X and $\overline{X}$.
c. Which distribution should you use for this problem?
d. Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the margin of error.
e. If the Census wants to increase its level of confidence and keep the margin of error the same by taking another survey, what changes should it make?
f. If the Census did another survey, kept the margin of error the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?
g. Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

---

5. A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

a. Identify the following:
   - $\overline{x}$
   - σ
   - n

b. In words, define the random variable X.
c. In words, define the random variable $\overline{X}$.

d. Which distribution should you use for this problem?
e. Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the margin of error.
f. Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the margin of error.
g. In complete sentences, explain why the confidence interval in (f) is larger than in (e).
h. In complete sentences, give an interpretation of what the interval in (e) means.
i. What would happen if 40 heads of lettuce were sampled instead of 20 and the margin of error remained the same?
j. What would happen if 40 heads of lettuce were sampled instead of 20 and the confidence level remained the same?

---

6. The mean age for all Foothill College students for a recent fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that 25 winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for winter Foothill College students.[10] Let X = the age of a winter Foothill College student.

a. Identify $\overline{x}$.
b. Identify n.
c. What variable does 15 represent?
d. In words, define the random variable $\overline{X}$.
e. What is $\overline{x}$ estimating?
f. Is σ_x known?
g. As a result of your answer to (e), state the exact distribution to use when calculating the confidence interval.
h. Construct a 95% confidence interval for the true mean age of winter Foothill College students by working out (i)-(o).
i. How much area is in both tails (combined)? α =______
j. How much area is in each tail? $\frac{\alpha}{2}$ =______
k. Identify the following specifications:

  - lower limit
  - upper limit
  - margin of error

l. The 95% confidence interval is:______.
m. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

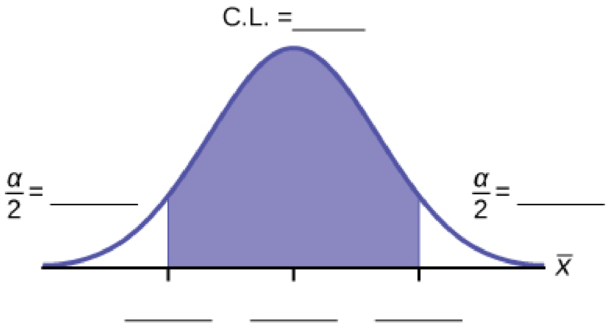


Figure 5.18. [Figure description available at the end of the section](#).

n. In one complete sentence, explain what the interval means.
o. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the margin of error become larger or smaller? How do you know?

---

7. Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

a. Identify $\overline{x}$.
b. Identify σ.
c. Identify n.
d. In words, define the random variables X and $\overline{X}$.
e. Which distribution should you use for this problem? Explain your choice.
f. Construct a 95% confidence interval for the population mean height of male Swedes.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

g. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

   - 71
   - 3
   - 48

8. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of *IEEE Spectrum* magazines. The mean length of the conferences was 3.94 days with a standard deviation of 1.28 days. Assume the underlying population is normal.

a. In words, define the random variables X and $\overline{X}$.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population mean length of engineering conferences.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

---

9. Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. Identify the following:

   - $\overline{x}$ = ______
   - σ = ______
   - n = ______

b. In words, define the random variables X and $\overline{X}$.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean time to complete the tax forms.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

e. If the firm wished to increase its level of confidence and keep the margin of error the same by taking another survey, what changes should it make?
f. If the firm did another survey, kept the margin of error the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

---

10. A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. Identify the following:
    - $\overline{x}$ = ______
    - σ = ______
    - s_x = ______
b. In words, define the random variable X.
c. In words, define the random variable $\overline{X}$.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 90% confidence interval for the population mean weight of the candies.
    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.
f. Construct a 98% confidence interval for the population mean weight of the candies.
    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.
g. In complete sentences, explain why the confidence interval in (f) is larger than the confidence interval in (e).
h. In complete sentences, give an interpretation of what the interval in f) means.

---

11. A camp director is interested in the mean number of letters each child sends during their camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

a. Identify the following:
    - $\overline{x}$ = ______
    - σ = ______
    - n = ______
b. Define the random variables X and $\overline{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean number of letters campers send home.
    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.

e. What will happen to the margin of error and confidence interval if 500 campers are surveyed? Why?

---

12. What is meant by the term “90% confident” when constructing a confidence interval for a mean?

a. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
b. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
c. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
d. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

---

13. The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. The figure below shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

Receipts				
$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

Figure 5.19

a. Find the point estimate for the population mean.
b. Using 95% confidence, calculate the margin of error.
c. Create a 95% confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

5.4 The Behavior of Confidence Intervals

1. Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of 6 minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 min.

The population standard deviation is six minutes, and the sample mean delivery time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

---

2. What happens to the margin of error in the pizza-delivery exercise if the sample size is changed? Leave everything the same except the sample size. Use the original 90% confidence level.

a. What happens to the margin of error and the confidence interval if we increase the sample size and use n = 100 instead of n = 36?
b. What happens if we decrease the sample size to n = 25 instead of n = 36?

- $\overline{x}$ = 68
- EBM = $\left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- σ = 3; the confidence level is 90% (CL = 0.90); $z_{\frac{\alpha}{2}}$ = $z_{0.05}$ = 1.645.

---

3. Refer back to the pizza-delivery exercise. The mean delivery time is 36 minutes, and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

5.5 Introduction to Hypothesis Tests

1. When do you reject the null hypothesis?

---

2. The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

3. It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The p-value is almost zero. State the null and alternative hypotheses, and interpret the p-value.

4. The mean age of graduate students at a university is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years, and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.

5. Does the shaded region represent a low or a high p-value compared to a level of significance of 1%?

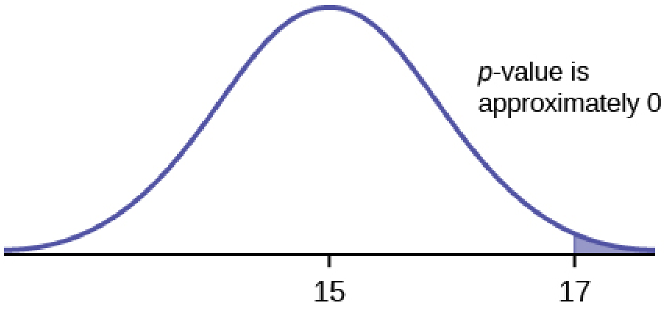


Figure 5.20. Figure description available at the end of the section.

6. What should you do when $\alpha > p$-value?

7. What should you do if $\alpha = p$-value?

8. *If you do not reject the null hypothesis, then it must be true.* Is this statement correct? State why or why not in complete sentences.

9. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

a. Is this a test of means or proportions?
b. What symbol represents the random variable for this test?
c. In words, define the random variable for this test.
d. Is σ known, and if so, what is it?
e. Calculate the following:

    - $\overline{x}$ ______
    - σ ______
    - s_x ______
    - n ______

f. Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.
g. State the distribution to use for the hypothesis test.

---

10. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: ______
   H_a : ______
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known, and if so, what is it?
g. Calculate the following:

    - $\overline{x}$ = ______
    - s = ______
    - n = ______

h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.

k. At a pre-conceived α = 0.05, what is your:
   - Decision
   - Reason for the decision
   - Conclusion (write out in a complete sentence)

---

11. The National Institute of Mental Health published an article stating that, in any one-year period, approximately 9.5% of American adults suffer from depression or a depressive illness.[11] Suppose that a survey of 100 people in a certain town found that seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: ______
   H_a: ______
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Calculate the following:
   - x = ______
   - n = ______
   - p' = ______
g. Calculate σ_x = ______. Show the formula set-up.
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived α = 0.05, what is your:
   - Decision
   - Reason for the decision
   - Conclusion (write out in a complete sentence)

---

12. We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ ______ 66
b. H_a: μ ______ 66

13. We want to test if college students take less than five years to graduate from college, on the average. State the null and alternative hypotheses.

---

14. We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ ______ 45
b. H_a: μ ______ 45

---

15. In an issue of *U.S. News & World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of US students take advanced placement exams and 4.4% pass. Test if the percentage of US students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

---

16. On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: p ______ 0.40
b. H_a: p ______ 0.40

---

17. You are testing that the mean speed of your cable Internet connection is more than three megabits per second. What is the random variable? Describe in words.

---

18. You are testing that the mean speed of your cable Internet connection is more than three megabits per second. State the null and alternative hypotheses.

---

19. The American family has an average of two children. What is the random variable? Describe in words.

---

20. The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

21. A sociologist claims the probability that a person picked at random in Times Square in New York City is a visitor to the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

---

22. A sociologist claims the probability that a person picked at random in Times Square in New York City is a visitor to the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

---

23. In a population of fish, approximately 42% are female. A test is conducted to see if the proportion is actually less. State the null and alternative hypotheses.

---

24. Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0: ______
b. H_a: ______

---

25. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0: ______
b. H_a: ______

---

26. The National Institute of Mental Health published an article stating that, in any one-year period, approximately 9.5% of American adults suffer from depression or a depressive illness.[12] Suppose that a survey of 100 people in a certain town found that 7 of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from

depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

a. H_0: ______
b. H_a: ______

---

27. Some of the following statements refer to the null hypothesis, some to the alternate hypothesis. State the null hypothesis, H_0, and the alternative hypothesis, H_a, in terms of the appropriate parameter (μ or p).

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

---

28. Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years, the girls were surveyed again. 63 said they smoked to stay thin. Is there good evidence that more than 30% of the teen girls smoke to stay thin? The alternative hypothesis is:

a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

---

29. A statistics instructor believes that fewer than 20% of Evergreen Valley College students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$

d. $p \leq 0.20$

30. Previously, an organization reported that teenagers spent an average of 4.5 hours on the phone each week. The organization thinks that the mean is currently higher. 15 randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

a. H_0: $\overline{x} = 4.5$; H_a: $\overline{x} > 4.5$
b. H_0: $\mu \geq 4.5$; H_a: $\mu < 4.5$
c. H_0: $\mu = 4.75$; H_a: $\mu > 4.75$
d. H_0: $\mu = 4.5$; H_a: $\mu > 4.5$

5.6 Hypothesis Tests in Depth

1. Suppose the null hypothesis, H_0, is: the victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

- Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive.
- Type II error: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.
- α = probability that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(type I error).
- β = probability that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(type II error).

Which is the error with the greater consequence?

2. Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, type I or type II?

3. It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy Genetic Labs has no effect on gender outcome.

- Type I error: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when it actually has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.
- Type II error: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

What is the error of greater consequence?

---

4. Red tide is a bloom of poison-producing algae composed of a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regularly sampling shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a type I and a type II error in this context, and state which error has the greater consequence.

---

5. A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the type I and type II errors in context. Which error is the more serious?

- Type I: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.
- Type II: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

---

6. Determine both type I and type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%.

---

7. Identify the type I and type II errors from these four statements:

a. Do not reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.
b. No not reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percent-

age is actually at least 88%.

d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

---

8. The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the type I and type II errors in complete sentences.

---

9. A sleeping bag is tested to withstand temperatures of –15°F. You think the bag cannot stand temperatures that low. State the type I and type II errors in complete sentences.

---

10. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences. Which is the error with the greater consequence?

---

11. The power of a test is 0.981. What is the probability of a type II error?

---

12. A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the type I and type II errors in complete sentences.

---

13. A microbiologist is testing a water sample for E. coli. Suppose the null hypothesis, H_0, is: the sample does not contain E. coli. The probability that the sample does not contain E. coli while the microbiologist thinks it does is 0.012. The probability that the sample does contain E. coli while the microbiologist thinks it does not is 0.002. What is the power of this test?

---

14. A microbiologist is testing a water sample for E. coli. Suppose the null hypothesis, H_0, is: the sample contains E. coli. Which is the error with the greater consequence?

15. State the type I and type II errors in complete sentences given the following statements.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in their lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

---

16. For statements (a) through (j) in the previous question, answer each of the following in complete sentences.

a. State a consequence of committing a type I error.
b. State a consequence of committing a type II error.

---

17. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration to market the drug. Suppose the null hypothesis is "the drug is unsafe." What is the type II error?

a. To conclude the drug is safe, when in fact, it is unsafe.
b. Not to conclude the drug is safe, when in fact, it is safe.
c. To conclude the drug is safe, when in fact, it is safe.
d. Not to conclude the drug is unsafe, when in fact, it is unsafe.

---

18. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The type I error is to conclude that the percent of EVC students who attended is ______.

a. At least 20%, when in fact, it is less than 20%
b. 20%, when in fact, it is 20%
c. Less than 20%, when in fact, it is at least 20%
d. Less than 20%, when in fact, it is less than 20%

19. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours.[13] At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven, when in fact, the mean number of hours is ______.

a. More than seven hours
b. At most seven hours
c. At least seven hours
d. Less than seven hours

---

20. Previously, an organization reported that teenagers spent an average of 4.5 hours on the phone each week. The organization thinks that the mean is currently higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The type I error is:

a. To conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher.
b. To conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same.
c. To conclude that the current mean hours per week is 4.5, when in fact, it is higher.
d. To conclude that the current mean hours per week is no higher than 4.5, when in fact, it is not higher.

Figure Descriptions

Figure 5.18: This is a template of a normal distribution curve with the central region shaded to represent a confidence interval. The residual areas are on either side of the shaded region. Blanks indicate that students should label the confidence level, residual areas, and points that define the confidence interval.

Figure 5.20: Normal distribution curve with one vertical upward line from x-axis to curve on the far right of the curve. From this line to the right is shaded under the curve. The mean is labeled with a value of 15, and the vertical line to the right of it is labeled with a value of 17. Above the curve reads ‘p-value is approximately 0.’

References

Figures

Figure 5.18: Figure 8.8 from OpenStax Introductory Statistics 2e (2023) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics-2e/pages/8-practice#element-158

Figure 5.20: Figure 9.18 from OpenStax Introductory Statistics 2e (2023) (CC BY 4.0). Retrieved from https://openstax.org/books/introductory-statistics-2e/pages/9-practice

Text

Baran, Daya. "20 Percent of Americans Have Never Used Email." WebGuild, 2010. Available online at http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at http://blog.flurry.com (accessed May 17, 2013).

Data from the United States Department of Agriculture.

Image credit: comedy_nose/flickr

"American Fact Finder." U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (accessed July 2, 2013).

"Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2, 2013).

"Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm (accessed September 30,2013).

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/growthcharts/2000growthchart-us.pdf (accessed July 2, 2013).

La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).

"Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table (accessed July 2, 2013).

"Metadata Description of Candidate Summary File." U.S. Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml (accessed July 2, 2013).

“National Health and Nutrition Examination Survey.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

(Credit: Robert Neff)

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Notes

1. La, Lynn, and Kent German. "Cell Phone Radiation Levels," CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels (accessed July 2, 2013).
2. “2012 College-Bound Seniors Total Group Profile Report,” CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).
3. Data from The Flurry Blog, 2013. Available online at http://blog.flurry.com (accessed May 17, 2013).
4. Data from The Flurry Blog, 2013. Available online at http://blog.flurry.com (accessed May 17, 2013).
5. Data from the United States Department of Agriculture.
6. “National Health and Nutrition Examination Survey,” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed May 17, 2013).
7. La, Lynn, and Kent German. "Cell Phone Radiation Levels," CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels (accessed July 2, 2013).
8. La, Lynn, and Kent German. "Cell Phone Radiation Levels," CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels (accessed July 2, 2013).
9. “American Fact Finder,” U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (accessed July 2, 2013).
10. “Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall,” Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends (accessed September 30,2013).
11. Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.
12. Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.
13. King, Bill. “Graphically Speaking,” Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Chapter 6 Extra Practice

6.1 The Sampling Distribution of the Sample Mean (t)

1. Which two distributions can you use for inference on a mean?

---

2. Which distribution do you use when you are testing a population mean and the population standard deviation is known and/or $n \geq 30$?

---

3. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

---

4. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

---

5. A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

---

7. You are performing a hypothesis test of a single population mean using a Student's t-distribution. What must you assume about the distribution of the data?

6.2 Inference for the Mean in Practice

1. The Human Toxome Project (HTP) is working to understand the scope of industrial pollution's effect on the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, in addition to fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. The figure below shows how many of the targeted chemicals were found in each infant's cord blood.[1]

Number of targeted chemicals									
79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Figure 6.8

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an infant's blood.

---

2. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variables X and $\overline{x}$ in words.
c. Which distribution should you use for this problem?
d. Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the margin of error.
e. Explain in complete sentences what the confidence interval means.

---

3. One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variable X in words.
c. Define the random variable $\overline{x}$ in words.
d. Which distribution should you use for this problem?

e. Construct a 99% confidence interval for the population mean hours spent watching television per month. State the confidence interval, sketch the graph, and calculate the margin of error.

f. Why would the margin of error change if the confidence level were lowered to 95%?

---

4. The data in the table below are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Frequency
1	1
2	7
3	18
4	7
5	6

Figure 6.9

a. Identify $\overline{x}$, s_x, and n.

b. Define the random variable $\overline{X}$ in words.

c. What is $\overline{X}$ estimating?

d. Is σ_x known?

e. As a result of your answer to the questions above, state the exact distribution to use when calculating the confidence interval.

f. Construct a 95% confidence interval for the true mean number of colors on national flags. How much area is in both tails (combined)?

g. How much area is in each tail?

h. Calculate the lower limit, upper limit, and margin of error.

i. The 95% confidence interval is ____.

j. Fill in the blanks on the graph with the areas, the upper and lower limits of the confidence interval, and the sample mean.

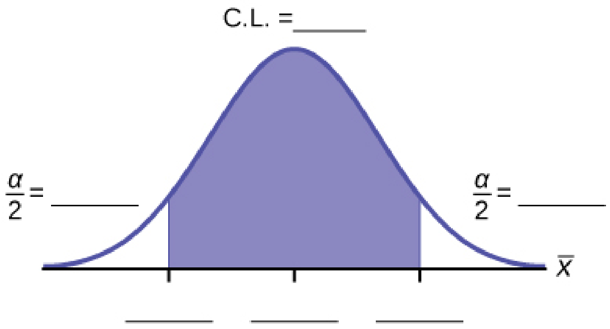


Figure 6.10. Figure description available at the end of the section.

k. In one complete sentence, explain what the interval means.
l. Using the same $\overline{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the margin of error become larger or smaller? How do you know?
m. Using the same $\overline{x}$, s_x, and n = 39, how would the margin of error change if the confidence level were reduced to 90%? Why?

---

6. A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variables X and $\overline{x}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

e. What would happen to the margin of error and confidence interval if 500 community colleges were surveyed? Why?

---

7. Suppose that a committee is studying whether or not there is time wasted in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variables X and $\overline{x}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean time wasted.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

e. Explain in a complete sentence what the confidence interval means.

8. A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7, 2.8, 3.0, 2.3, 2.3, 2.2, 2.8, 2.1, and 2.4.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variable X in words.
c. Define the random variable $\overline{x}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 95% confidence interval for the population mean length of time.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

f. What does it mean to be "95% confident" in this problem?

---

9. Suppose that 14 children who were learning to ride two-wheel bikes were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variable X in words.
c. Define the random variable $\overline{x}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 99% confidence interval for the population mean length of time using training wheels.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the margin of error.
f. Why would the margin of error change if the confidence level were lowered to 90%?

---

10. *Forbes* magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The figure below shows the ages of the corporate CEOs for a random sample of these firms.[2]

Ages of the corporate CEOs				
48	58	51	61	56
59	74	63	53	50

Ages of the corporate CEOs				
59	60	60	57	55
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Figure 6.11

Use this sample data to construct a 90% confidence interval for the mean age of CEOs for these top small firms. Use the Student's *t*-distribution.

---

11. Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats, and the sample standard deviation is 4.1 seats.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variables X and $\overline{x}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.

    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.

---

12. In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable $\overline{x}$ in words.
c. Construct a 95% confidence interval for the population mean cost of a used car.

    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.

d. Explain what a "95% confidence interval" means for this study.

13. Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are 8, 8, 10, 7, 9, and 9. Assume the underlying distribution is approximately normal.

a. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

b. If you wanted a smaller margin of error while keeping the same level of confidence, what should have been changed in the study before it was done?
c. Go to the store, and record the grams of fat per serving of six brands of chocolate chip cookies.
d. Calculate the mean.
e. Is the mean within the interval you calculated in (a)? Did you expect it to be? Why or why not?

---

14. A survey of the mean number of cents saved from coupons was conducted by randomly surveying one coupon per page from the coupon sections of a recent *San Jose Mercury News*. The following data were collected: 20¢, 75¢, 50¢, 65¢, 30¢, 55¢, 40¢, 40¢, 30¢, 55¢, $1.50, 40¢, 65¢, 40¢.[3] Assume the underlying distribution is approximately normal.

a. Identify $\overline{x}$, s_x, n, and n – 1.
b. Define the random variables X and $\overline{x}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean worth of coupons.

   - State the confidence interval.
   - Sketch the graph.
   - Calculate the margin of error.

e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

---

15. A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Find the 95% confidence Interval for the true population mean for the amount of soda served.

a. Find the 95% confidence Interval for the true population mean for the amount of soda served.
b. What is the margin of error?

---

12. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\overline{X}$ ~ ____.

a. N(7.24, $\frac{1.93}{\sqrt{22}}$)
b. N(7.24, 1.93)
c. t_{22}
d. t_{21}

6.3 The Sampling Distribution of the Sample Proportion

1. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of *np* and *nq*?

---

2. You are performing a hypothesis test of a single population proportion. You find out that *np* is less than five. What must you do to be able to perform a valid hypothesis test?

---

3. You are performing a hypothesis test of a single population proportion. The data come from which distribution?

6.4 Inference for a Proportion

1. In six packages of The Flintstones® Real Fruit Snacks, there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

a. Define the random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem? Explain your choice
c. Calculate $\hat{p}$.
d. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.

  - State the confidence interval.
  - Sketch the graph.
  - Calculate the margin of error.

e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

---

2. For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

---

3. A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.
b. In a sample of 300 students, 68% said they own an iPod and a smartphone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

---

4. Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?
b. If it were later determined that it was important to be more than 90% confident and a new survey were

commissioned, how would it affect the minimum number you need to survey? Why?

---

5. Suppose the marketing company randomly surveyed 200 households and found that, in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

a. Identify x, n, and $\hat{p}$.
b. Define the random variables X and $\hat{p}$ in words.
c. Which distribution should you use for this problem?
d. Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the margin of error.
e. List two difficulties the company might have in obtaining random results if this survey were done by email.

---

6. Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

a. We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem?
c. Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the margin of error.
d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given in the survey is ±2%. Explain what the ±2% means.

---

7. A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

a. Define the random variable X in words.
b. Define the random variable $\hat{p}$ in words.
c. Which distribution should you use for this problem?
d. Construct a 90% confidence interval, and state the confidence interval and the margin of error.
e. What would happen to the confidence interval if the level of confidence were 95%?

8. The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 PM Monday night beginners class for ages 8 to 12 was picked. In that class, were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls ages 8 to 12 in all beginners ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

a. What is being counted?
b. In words, define the random variable X.
c. Calculate x, n, and $\hat{p}$.
d. State the estimated distribution of X. X~ _____.
e. Define a new random variable $\hat{p}$. What is $\hat{p}$ estimating?
f. In words, define the random variable $\hat{p}$.
g. State the estimated distribution of $\hat{p}$. Construct a 92% confidence interval for the true proportion of girls in the ages 8 to 12 beginners ice-skating classes at the Ice Chalet.
h. How much area is in both tails (combined)?
i. How much area is in each tail?
j. Calculate the lower limit, upper limit, and margin of error.
k. The 92% confidence interval is ______.
l. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

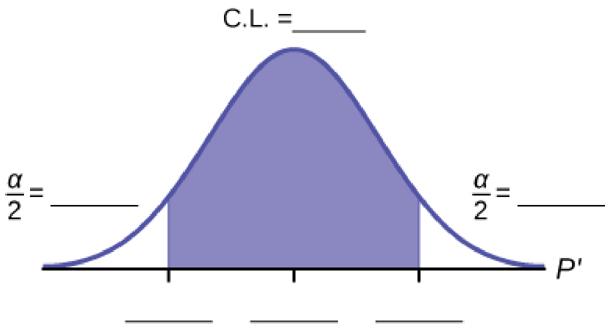


Figure 6.12. Figure description available at the end of the section.

m. In one complete sentence, explain what the interval means.
n. Using the same $\hat{p}$ and level of confidence, suppose that n were increased to 100. Would the margin of error become larger or smaller? How do you know?
o. Using the same $\hat{p}$ and n = 80, how would the margin of error change if the confidence level were increased to 98%? Why?
p. If you decreased the allowable margin of error, why would the minimum sample size increase (keeping the same level of confidence)?

9. Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

a. Find x, n, and $\hat{p}$.
b. Define the random variables X and $\hat{p}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population proportion who claim they always buckle up.

    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.

e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

---

10. According to a survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

a. Define the random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.

    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.

---

11. An article regarding interracial dating and marriage appeared in the *Washington Post*. Of the 1,709 randomly selected adults, 315 identified themselves as Latino/a, 323 identified themselves as Black, 254 identified themselves as Asian, and 779 identified themselves as White. In this survey, 86% of Black people said that

they would welcome a White person into their families. Among Asian people, 77% would welcome a White person into their families, 71% would welcome a Latino/a person, and 66% would welcome a Black person.[4]

a. We are interested in finding the 95% confidence interval for the percent of all Black adults who would welcome a White person into their families. Define the random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval.
    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.

---

12. Refer to the information in Question 11.

a. Construct three 95% confidence intervals.
    - Percent of all Asian people who would welcome a White person into their families
    - Percent of all Asian people who would welcome a Latino/a into their families
    - Percent of all Asian people who would welcome a Black person into their families
b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

---

13. Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

a. Define the random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.
d. Explain what a "97% confidence interval" means for this study.

14. A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.”[5] We are interested in the population proportion of adult Americans who feel that crime is the main problem.

a. Define the random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.
d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

---

15. Refer to the information above. Another question in the poll was “[How much are] you worried about the quality of education in our schools?” 63% percent responded “a lot.” We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

a. Define the random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
    - State the confidence interval.
    - Sketch the graph.
    - Calculate the margin of error.
d. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one-to-three complete sentences, explain what the ±3% represents.

---

16. According to a Field Poll, 79% of California adults (400 out of 506 surveyed) feel that “education and our schools” is one of the top issues facing California.[6] We wish to construct a 90% confidence interval for the true proportion of California adults who feel that “education and the schools” is one of the top issues facing California.

a. What is a point estimate for the true population proportion?
b. A 90% confidence interval for the population proportion is ______.

c. The margin of error is approximately _______.

---

17. In a certain Southern California community, 511 homes are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. Of the homes surveyed, 173 met the minimum recommendations for earthquake preparedness, and 338 did not.

a. Find the confidence interval at the 90% confidence level for the true population proportion of Southern California community homes meeting at least the minimum recommendations for earthquake preparedness.
b. The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

---

18. On May 23, 2013, Gallup reported that, of the 1,005 people surveyed, 76% of US workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a ±3% margin of error.[7]

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify CL and α.
d. Calculate the margin of error based on the information provided.
e. Compare the margin of error in (d) to the margin of error reported by Gallup. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

---

19. A national survey of 1,000 adults was conducted on May 13, 2013, by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.[8]

a. Find the point estimate and the margin of error for this confidence interval.
b. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from (a) and n = 1,000 to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

20. Public Policy Polling recently conducted a survey asking adults across the US about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.[9]

a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The margin of error of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

---

21. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2020 presidential election with 95% confidence and a margin of error no greater than 5%. How many students must you interview?

---

22. In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their community as "likely" or "very likely."[10] Use the "plus four" method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

6.5 Behavior of Confidence Intervals for a Proportion

1. The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users.[11] In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the "plus four" method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Figure Descriptions

Figure 6.10: This is a template of a normal distribution curve with the central region shaded to represent a confidence interval. The residual areas are on either side of the shaded region. Blanks indicate that students should label the confidence level, residual areas, and points that define the confidence interval.

Figure 6.12: This is a template of a normal distribution curve with the central region shaded to represent a confidence interval. The residual areas are on either side of the shaded region. Blanks indicate that students should label the confidence level, residual areas, and points that define the confidence interval.

References

Figures

Figure 6.10: Figure from Lumen Learning Introduction to Statistics (CC BY 4.0). Retrieved from https://courses.lumenlearning.com/introstats1/chapter/section-exercises-7/

Figure 6.12: Figure from Lumen Learning Introduction to Statistics (CC BY 4.0). Retrieved from https://courses.lumenlearning.com/introstats1/chapter/section-exercises-7/

Text

"America's Best Small Companies," *Forbes*, 2013. Available online at http://www.forbes.com/best-small-companies/list (accessed July 2, 2013).

Data from *Microsoft Bookshelf*.

Data from *Business Week*. Available online at http://www.businessweek.com.

Data from *Forbes*. Available online at http://www.forbes.com.

"Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012," Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2,2013).

"Human Toxome Project: Mapping the Pollution in People," Environmental Working Group. Available online at http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn (accessed July 2, 2013).

"Metadata Description of Leadership PAC List," Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml (accessed July 2, 2013).

Jensen, Tom. "Democrats, Republicans Divided on Opinion of Music Icons," Public Policy Polling. Available online at http://www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).

Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. "Teens, Social Media, and Privacy," PewInternet, 2013. Available online at http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx (accessed July 2, 2013).

Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey,” Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

Saad, Lydia. “Three in Four U.S. Workers Plan to Work Past Retirement Age: Slightly more say they will do this by choice rather than necessity,” *Gallup Economy*, 2013. Available online at http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx (accessed July 2, 2013).

The Field Poll. Available online at http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).

“New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security,” Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll (accessed July 2, 2013).

“52% Say Big-Time College Athletics Corrupt Education Process,” Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).

Notes

1. “Human Toxome Project: Mapping the Pollution in People,” Environmental Working Group. Available online at http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn (accessed July 2, 2013).
2. “America’s Best Small Companies.” Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).
3. Data from *San Jose Mercury News*
4. Fears, Darryl, and Claudia Deane. "Biracial Couples Report Tolerance," *Washington Post*, July 5, 2001. Available online at https://www.washingtonpost.com/archive/politics/2001/07/05/biracial-couples-report-tolerance/c1ce88c8-ba7c-44f5-a348-b86776df9112 (accessed January 26, 2021).
5. Data from *Time Magazine*; survey by Yankelovich Partners, Inc.
6. The Field Poll. Available online at http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).
7. Saad, Lydia. “Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity,” *Gallup Economy*, 2013. Available online at http://www.gallup.com/poll/162758/three-fourworkers-plan-work-past-retirement-age.aspx (accessed July 2, 2013).
8. “52% Say Big-Time College Athletics Corrupt Education Process,” Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).
9. Jensen, Tom. “Democrats, Republicans Divided on Opinion of Music Icons,” Public Policy Polling. Available online at http://www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).
10. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security,” Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/299-americans-neither-worried-norprepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll (accessed July 2, 2013).
11. Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey,” Pew Research Center:

Internet and American Life Project. Available online at http://www.pewinternet.org/~/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

Chapter 7 Extra Practice

7.1 Inference for Two Dependent Samples (Matched Pairs)

1. Seven eighth graders at Kennedy Middle School measured how far (measured in feet) they could throw the shot put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could throw equal distances with either hand. The data were collected and recorded below.

Hand used	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant hand	30	26	34	17	19	26	20
Weaker hand	28	14	27	18	17	26	16

Figure 7.9

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant and weaker hands is significant.

---

2. Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded below. Conduct a hypothesis test to determine whether the mean difference in distances between the dominant and off-hand is significant. Test at the 5% level.

	Player 1	Player 2	Player 3	Player 4	Player 5
Dominant hand	120	111	135	140	125
Off-hand	105	109	98	111	99

Figure 7.10

---

3. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown below. The "before" value is matched to an "after" value, and the differences are calculated. The differences have a normal distribution. Test at the 1% significance level.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

Figure 7.11

a. What is the random variable?
b. State the null and alternative hypotheses.
c. What is the *p*-value?
d. Draw the graph of the *p*-value.
e. What conclusion can you draw about the software patch?

---

4. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution. Test at the 1% significance level.

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

Figure 7.12

a. State the null and alternative hypotheses.
b. What is the *p*-value?
c. What is the sample mean difference?
d. Draw the graph of the *p*-value.
e. What conclusion can you draw about the juggling class?

---

5. A doctor wants to know if a blood pressure medication is effective. Six subjects have their systolic blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. Test at the 1% significance level.

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

Figure 7.13

a. State the null and alternative hypotheses.
b. What is the test statistic?
c. What is the *p*-value?
d. What is the sample mean difference?
e. What is the conclusion?

NOTE: *If you are using a Student's t-distribution, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)*

6. Ten individuals went on a low–fat diet for 12 weeks to lower their cholesterol. The data are recorded below. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
280	150
290	200
360	300
280	300
260	240

Figure 7.14

---

7. A new AIDS prevention drug was tried on a group of 224 HIV-positive patients. Forty-five patients developed AIDS after four years. In a control group of 224 HIV-positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript *t* = treated patient and *ut* = untreated patient.

The appropriate hypotheses are:

a. H_0: $p_t < p_{ut}$ and H_a: $p_t \geq p_{ut}$
b. H_0: $p_t \leq p_{ut}$ and H_a: $p_t > p_{ut}$
c. H_0: $p_t = p_{ut}$ and H_a: $p_t \neq p_{ut}$

d. H_0: $p_t = p_{ut}$ and H_a: $p_t < p_{ut}$

If the p-value is 0.0062 what is the conclusion? Use $\alpha = 0.05$.

a. The method has no effect.
b. There is sufficient evidence to conclude that the method reduces the proportion of HIV-positive patients who develop AIDS after four years.
c. There is sufficient evidence to conclude that the method increases the proportion of HIV-positive patients who develop AIDS after four years.
d. There is insufficient evidence to conclude that the method reduces the proportion of HIV-positive patients who develop AIDS after four years.

---

8. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a "biofeedback exercise program." Six subjects were randomly selected, and blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after – before) producing the following results: $\overline{x}_d = -10.2$ and $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

a. What is the distribution for the test?
b. If $\alpha = 0.05$, what is the p-value and the conclusion?

---

9. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She records the 18-hole scores for four new students before they learned the technique and then after they took her class. She conducts a hypothesis test. The data are as follows.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Figure 7.15

The correct decision is:

a. Reject H_0.
b. Do not reject H_0.

10. A local cancer support group believes that the estimate for new female breast cancer cases in the South is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are shown below.

Southern states	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

Figure 7.16

Test: Two matched pairs or paired samples (*t*-test)

Random variable: $\overline{X}_d$

Distribution: t_{12}

H_0: $\mu_d = 0$

H_a: $\mu_d > 0$

The mean of the differences of new female breast cancer cases in the South between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the South is higher in 2013 than in 2012.

Graph: Right-tailed

p-value: 0.0004

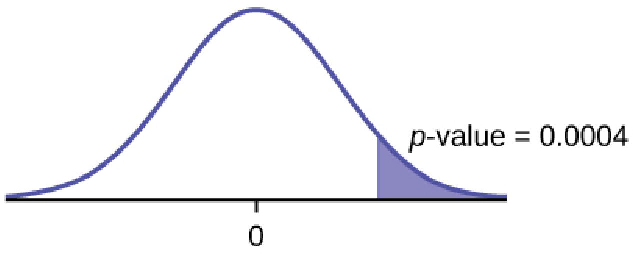


Figure 7.17. Figure description available at the end of the section.

Decision: Reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

---

11. A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is shown below. Test at the 1% level of significance.

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	469
Philadelphia	179	159
Washington, DC	245	239

Figure 7.18

---

12. A politician asked his staff to determine whether the underemployment rate in the Northeast decreased from 2011 to 2012. The results are shown in Figure 7.19.

Northeastern states	2012	2013
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

Figure 7.19

Test: Matched or paired samples (*t*-test)

Difference data: {−0.9, −3.7, −3.2, −0.5, 0.6, −1.9, −0.5, 0.2, 0.6, 0.4, 1.7, −2.4, 1.8}

Random variable: $\overline{X}_d$

Distribution:

H_0: $\mu_d = 0$

H_a: $\mu_d < 0$

The mean of the differences of the rate of underemployment in the Northeastern states between 2012 and 2011 is less than zero. The underemployment rate went down from 2011 to 2012.

Graph: Left-tailed

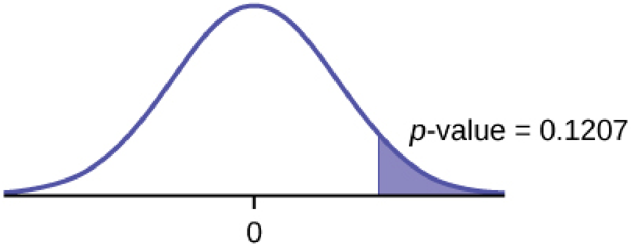


Figure 7.20. *Figure description available at the end of the section*.

p-value: 0.1207

Decision: Do not reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the Northeastern states from 2011 to 2012.

For Questions 13–22, indicate which of the following choices best identifies the hypothesis test:

a. *independent group means, population standard deviations and/or variances known*
b. *independent group means, population standard deviations and/or variances unknown*
c. *matched or paired samples*
d. *single mean*
e. *two proportions*
f. *single proportion*

13. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The population standard deviations are two pounds and three pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

---

14. A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children who like the new chocolate bar is greater than the proportion of adults who like it.

---

15. The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 9 male students and 16 female students.

---

16. A football league reported that the mean number of touchdowns per game was five. A study is done to determine if the mean number of touchdowns has decreased.

---

17. A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and one year, respectively.

18. According to a YWCA Rape Crisis Center newsletter, 75% of rape victims know their attackers. A study is done to verify this.

19. According to a recent study, US companies have a mean maternity leave of six weeks.

20. A recent drug survey showed an increase in the use of drugs and alcohol among local high school students as compared to the national percent. Suppose that a survey of 100 local youths and 100 national youths is conducted to see if the proportion of drug and alcohol use is higher locally than nationally.

21. A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Pre-course score	Post-course score
1	300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200
740	850

Figure 7.21

22. University of Michigan researchers reported in the *Journal of the National Cancer Institute* that quitting smoking is especially beneficial for those under age 49. In this American Cancer Society study, the risk (probability) of dying of lung cancer was about the same as for those who had never smoked.[1]

23. Lesley E. Tan investigated the relationship between left-handedness vs. right-handedness and motor competence in preschool children. Random samples of 41 left-handed preschool children and 41 right-handed preschool children were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The experiment produced the means and standard deviations shown below. Determine the appropriate test and best distribution to use for that test.

	Left-handed	Right-handed
Sample size	41	41
Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

Figure 7.22

a. Two independent means, normal distribution
b. Two independent means, Student's *t*-distribution
c. Matched or paired samples, Student's *t*-distribution
d. Two population proportions, normal distribution

---

24. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She records the 18-hole scores of four new students before they learned the technique and then after they took her class. She conducts a hypothesis test. The data are:

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Figure 7.23

This is:

a. a test of two independent means.
b. a test of two proportions.
c. a test of a single mean.
d. a test of a single proportion.

7.2 Inference for Two Independent Sample Means

1. The mean lasting time of two competing floor waxes is to be compared. Twenty floors are randomly assigned to test each wax. Both populations have a normal distributions. The data are recorded below. Does the data indicate that Wax 1 is more effective than Wax 2? Test at a 5% level of significance.

Wax	Sample mean number of months floor wax lasts	Population standard deviation
1	3	0.33
2	2.9	0.36

Figure 7.24

---

2. The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. The data are recorded below. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5% level of significance.

Engine	Sample mean number of RPM	Population standard deviation
1	1,500	50
2	1,600	60

Figure 7.25

---

3. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. The data are recorded below. Scouters believe that Rodriguez pitches a speedier fastball.

Pitcher	Sample mean speed of pitches (mph)	Population standard deviation
Wesley	86	3
Rodriguez	91	7

Figure 7.26

a. What is the random variable?
b. State the null and alternative hypotheses.
c. What is the test statistic?
d. What is the p-value?
e. At the 1% significance level, what is your conclusion?

4. A researcher is testing the effects of plant food on plant growth. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of the plants are recorded after eight weeks. The populations have normal distributions. The following table is the result. The researcher thinks the food makes the plants grow taller.

Plant group	Sample mean height of plants (inches)	Population standard deviation
Food	16	2.5
No food	14	1.5

Figure 7.27

a. Is the population standard deviation known or unknown?
b. State the null and alternative hypotheses.
c. What is the p-value?
d. Draw the graph of the p-value.
e. At the 1% significance level, what is your conclusion?

---

5. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. Fifteen pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point.

	Sample mean melting temperatures (°F)	Population standard deviation
Alloy Gamma	800	95
Alloy Zeta	900	105

Figure 7.28

a. State the null and alternative hypotheses.
b. Is this a right-, left-, or two-tailed test?
c. What is the p-value?
d. Draw the graph of the p-value.
e. At the 1% significance level, what is your conclusion?

---

6. Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

Subscripts: 1 = boys, 2 = girls

a. State the null and alternative hypotheses.
b. What is the random variable?
c. Which distribution should you use for this problem?
d. What is the test statistic?
e. What is the p-value?
f. At the 5% significance level, what is your decision and conclusion?

---

7. A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were $947 and $1,011, respectively. The population standard deviations are known to be $254 and $87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

---

8. Some manufacturers claim that non-hybrid sedan cars have a lower mean miles per gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of seven mpg. Thirty-one non-hybrid sedans get a mean of 22 mpg with a standard deviation of four mpg. Suppose that the population standard deviations are known to be six and three, respectively. Conduct a hypothesis test to evaluate the manufacturers claim.

Subscripts: 1 = non-hybrid sedans, 2 = hybrid sedans

a. State the null and alternative hypotheses.
b. What is the random variable?
c. Which distribution should you use for this problem?
d. What is the test statistic?
e. What is the p-value?
f. At the 5% significance level, what is your decision and conclusion?

---

9. A baseball fan wanted to know if there is a difference between the number of games played in a World Series when the American League won the series versus when the National League won the series. From 1922 to 2012, the population standard deviation of games won by the American League was 1.14, and the population standard deviation of games won by the National League was 1.11. Of 19 randomly selected World Series games won by the American League, the mean number of games won was 5.76. The mean number of 17 randomly selected games won by the National League was 5.42. Conduct a hypothesis test.

10. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from one (strongly agree) to five (strongly disagree). The table below contains ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Husband's score	Wife's score
2	2
2	2
1	3
3	3
2	4
1	2
1	1
1	1
2	2
4	4

Figure 7.29

11. The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data below. Each populations has a normal distribution.

	Sample size	Average number of hours playing sports per day	Sample standard deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Figure 7.30

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

12. Two samples are shown in the table below. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5% level of significance.

	Sample size	Sample mean	Sample standard deviation
Population A	25	5	1
Population B	16	4.7	1.2

Figure 7.31

NOTE: When the sum of the sample sizes is larger than 30 ($n_1 + n_2 > 30$), you can use the normal distribution to approximate the Student's *t*.

13. A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from College A has taken more math classes, on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

a. Is this a test of two means or two proportions?
b. Are the populations standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternate hypotheses? Write the null and alternate hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the p-value?
h. Do you reject or not reject the null hypothesis?
i. What is the conclusion?

14. A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

a. Are the population standard deviations known?
b. Conduct an appropriate hypothesis test. At the 5% significance level, what is your conclusion?

15. A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed below.

Online class: 67.6, 41.2, 85.3, 55.9, 82.4, 91.2, 73.5, 94.1, 64.7, 64.7, 70.6, 38.2, 61.8, 88.2, 70.6, 58.8, 91.2, 73.5, 82.4, 35.5, 94.1, 88.2, 64.7, 55.9, 88.2, 97.1, 85.3, 61.8, 79.4, 79.4

Face-to-face class: 77.9, 95.3, 81.2, 74.1, 98.8, 88.2, 85.9, 92.9, 87.1, 88.2, 69.4, 57.6, 69.4, 67.1, 97.6, 85.9, 88.2, 91.8, 78.8, 71.8, 98.8, 61.2, 92.9, 90.6, 97.6, 100, 95.3, 83.5, 92.9, 89.4

Is the mean of the final exam scores of the online class lower than the mean of the final exam scores of the face-to-face class? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right, left, or two tailed?
g. What is the p-value?
h. Do you reject or not reject the null hypothesis?
i. At the _____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

---

16. Cohen's d is a measure of effect size based on the differences between two means. Cohen's d, named for United States statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Size of effect	*d*
Small	0.2
Medium	0.5
Large	0.8

Figure 7.32

Cohen's d is the measure of the difference between two means divided by the pooled standard deviation,

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}}$$

where

$$s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}.$$

Calculate Cohen's *d* for Question 14. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

---

17. Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen, while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the Northeast and in the West as identified by Nasdaq on May 24, 2013, are listed in the tables below.

Northeast: 94.2, 75.2, 69.6, 52.0, 48.0, 41.9, 36.4, 33.4, 31.5, 27.6, 77.3, 71.9, 67.5, 50.6, 46.2, 38.4, 35.2, 33.0, 28.7, 26.5, 76.3, 71.7, 56.3, 48.7, 43.2, 37.6, 33.7, 31.8, 28.5, 26.0

West: 126.0, 70.6, 65.2, 51.4, 45.5, 37.0, 33.0, 29.6, 23.7, 22.6, 116.1, 70.6, 58.2, 51.2, 43.2, 36.0, 31.4, 28.7, 23.5, 21.6, 78.2, 68.2, 55.6, 50.3, 39.0, 34.1, 31.0, 25.3, 23.4, 21.5

Is there a difference in the weighted alpha of the top 30 stocks of banks in the Northeast and in the West? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the *p*-value?
h. Do you reject or not reject the null hypothesis?
i. At the ____ level of significance, from the sample data, there ____ (is/is not) sufficient evidence to conclude that ____.
j. Calculate Cohen's *d* and interpret it.

For the next 15 exercises, indicate whether the hypothesis test is for:

a. *independent group means, population standard deviations, and/or variances known*
b. *independent group means, population standard deviations, and/or variances unknown*
c. *matched or paired samples*
d. *single mean*
e. *two proportions*

f. *single proportion*

18. It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. Of interest is whether the proportions are in fact equal.

19. A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

20. A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

21. The known standard deviation in salary for all mid-level professionals in the financial industry is $11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is $80,000. The sample mean salary for mid-level professionals in Company B is $96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.

22. The average worker in Germany gets eight weeks of paid vacation.

23. According to a television commercial, 80% of dentists agree that Ultrafresh toothpaste is the best on the market.

24. It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.

25. The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the Vikings. Are the batting averages of the Rattlers and Vikings statistically different?

26. In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

27. A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after.

28. It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

29. Varsity athletes practice five times a week, on average.

30. A sample of 12 in-state graduate school programs at School A has a mean tuition of $64,000 with a standard deviation of $8,000. At School B, a sample of 16 in-state graduate programs has a mean of $80,000 with a standard deviation of $6,000. On average, are the mean tuitions different?

31. A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

32. A high school principal claims that 30% of student athletes drive themselves to school, while 4% of non-athletes drive themselves to school. In a sample of 20 student athletes, 45% drive themselves to school. In a sample of 35 non-athlete students, 6% drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

33. A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

a. Are standard deviations known or unknown?
b. What is the random variable?
c. The random variable is the difference between the mean amounts of sugar in the two soft drinks.
d. Is this a one-tailed or two-tailed test?

---

34. The US Center for Disease Control reports that the mean life expectancy was 47.6 years for White people born in 1900 and 33.0 years for non-White people. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 White people, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 non-White people, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for White people and non-White people.

a. Is this a test of means or proportions?
b. State the null and alternative hypotheses.
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. In symbols, what is the random variable of interest for this test?
e. In words, define the random variable of interest for this test.
f. Which distribution (normal or Student's t) would you use for this hypothesis test?
g. Explain why you chose the distribution you did.
h. Calculate the test statistic and p-value.
i. Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p-value.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, what is your:

  - Decision
  - Reason for the decision
  - Conclusion (write out in a complete sentence)

l. Does it appear that the means are the same? Why or why not?

NOTE: *If you are using a Student's t-distribution, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)*

35. The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted, and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same?

36. A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

Subscripts: 1: two-year colleges, 2: four-year colleges

37. At Rachel's 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Figure 7.33

38. Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry-level mechanical engineers and 60 entry-level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3,450 and $4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

Subscripts: 1: mechanical engineering, 2: electrical engineering

---

39. Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

---

40. The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals. Conduct a hypothesis test. What is the exact distribution for the hypothesis test? If the level of significance is 0.05, what is the conclusion?

Western	Eastern
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Figure 7.34

---

41. Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. A concluding statement is:

a. There is sufficient evidence to conclude that the statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
b. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
c. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
d. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

42. Researchers interviewed street sex workers in Canada and the United States. The mean age of the 100 Canadian sex workers upon entering sex work was 18 with a standard deviation of six. The mean age of the 130 United States sex workers upon entering their work was 20 with a standard deviation of eight. Is the mean age of entering sex work in Canada lower than the mean age in the United States? Test at a 1% significance level.

---

43. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds. Conduct a hypothesis test.

---

44. Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The "day" subscript refers to the statistics day students. The "night" subscript refers to the statistics night students. An appropriate alternative hypothesis for the hypothesis test is:

a. $\mu_{day} > \mu_{night}$
b. $\mu_{day} < \mu_{night}$
c. $\mu_{day} = \mu_{night}$
d. $\mu_{day} \neq \mu_{night}$

7.3 Inference for Two-Sample Proportions

1. A research study was conducted about gender differences in "sexting." The researcher believed that the proportion of girls involved in sexting is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized below. Is the proportion of girls sending sexts less than the proportion of boys sexting? Test at a 1% level of significance.[2]

	Males	Females
Sent sexts	183	156
Total number surveyed	2,231	2,169

Figure 7.35

2. Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with White, non-Hispanic people than with African American people. The results of the survey indicate that, of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 White cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of White iPhone owners greater than the proportion of African American iPhone owners?[3]

---

3. An interested citizen wanted to know if Democratic US senators are older than Republican US senators, on average. On May 26 2013, the mean age of 30 randomly selected Republican Senators was 61 years 247 days old (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days old (61.704 years) with a standard deviation of 9.55 years.[4]

Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5% level of significance.

---

4. A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that, of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category.[5] Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Which distribution do you use to perform the test?
c. What is the random variable?
d. What are the null and alternative hypothesis? Write the null and alternative hypothesis in symbols.
e. Is this test right-, left-, or two-tailed?
f. What is the p-value?
g. Do you reject or not reject the null hypothesis?
h. At the ____ level of significance, from the sample data, there ____ (is/is not) sufficient evidence to conclude that ____.

---

5. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1.

a. Is this a test of means or proportions?
b. What is the random variable?

c. State the null and alternative hypotheses.
d. What is the p-value?
e. What can you conclude about the two operating systems?

---

6. In the recent Census, three percent of the US population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nine people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races.[6] Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

a. Is this a test of means or proportions?
b. State the null and alternative hypotheses.
c. Is this a right-tailed, left-tailed, or two-tailed test? How do you know?
d. What is the random variable of interest for this test?
e. In words, define the random variable for this test.
f. Which distribution (normal or Student's t) would you use for this hypothesis test?
g. Explain why you chose the distribution you did.
h. Calculate the test statistic.
i. Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p-value.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, what is your:

   - Decision
   - Reason for the decision
   - Conclusion (write out in a complete sentence)

l. Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

---

7. A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them.

Indicate which of the following choices best identifies the hypothesis test.

a. independent group means, population standard deviations and/or variances known
b. independent group means, population standard deviations and/or variances unknown
c. matched or paired samples
d. single mean
e. two proportions
f. single proportion

---

8. A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from the California state university system and 100 from private universities are surveyed. Suppose that, from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

Indicate which of the following choices best identifies the hypothesis test.

a. independent group means, population standard deviations and/or variances known
b. independent group means, population standard deviations and/or variances unknown
c. matched or paired samples
d. single mean
e. two proportions
f. single proportion

---

9. We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for the White and Black races in the United States. We randomly pick one year, 1992, to compare the races. The number of suicides estimated in the United States in 1992 for White females is 4,930, of whom 580 were aged 15 to 24. The estimate for Black females is 330, of whom 40 were aged 15 to 24.[7] We will let female suicide victims be our population.

---

10. Elizabeth Mjelde, an art history professor, was interested in whether the value from the golden ratio formula $\left(\frac{\text{larger + smaller dimension}}{\text{larger dimension}}\right)$ was the same in the Whitney Exhibit for works from 1900 to 1919 as for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064.[8] Do you think that there is a significant difference in the golden ratio calculation?

11. A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students.[9] In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?

Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College

---

12. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system. It is spread by the Culex species of mosquito. In the United States in 2010, there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases, and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011.[10] Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test.

Subscripts: 2011 = 2011 group, 2010 = 2010 group

a. This is:

a. a test of two proportions.
b. a test of two independent means.
c. a test of a single mean.
d. a test of matched pairs.

b. An appropriate null hypothesis is:

a. $p_{2011} \leq p_{2010}$.
b. $p_{2011} \geq p_{2010}$.
c. $\mu_{2011} \leq \mu_{2010}$.
d. $p_{2011} > p_{2010}$.

c. The p-value is 0.0022. At a 1% level of significance, the appropriate conclusion is

a. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
b. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
c. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
d. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States

in 2010 who contracted neuroinvasive West Nile disease.

---

13. Researchers conducted a study to find out if there is a difference in the use of eReaders by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7% of the 628 surveyed use eReaders, while 11% of the 2,309 participants 30 years old and older use eReaders.[11]

---

14. Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the South is less than the proportion of southern men who are obese. The results are shown below.[12] Test at the 1% level of significance.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

Figure 7.36

---

15. Two computer users were discussing tablet computers. At one point in time, a higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. The figure below details the number of tablet owners for each age group. Test at the 1% level of significance.

	16–29 years old	30 years old and older
Own a tablet	69	231
Sample size	628	2,309

Figure 7.37

---

16. A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that, of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones.[13] Test at the 5% level of significance.

17. While her husband spent 2.5 hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

Subscripts: 1: men, 2: women

---

18. We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

---

19. Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of males has reached the proportion of females?

---

20. "To Breakfast or Not to Breakfast?" by Richard Ayore:

In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, …, birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, "This is why I always work more hours than you guys!"

And so, to prove him wrong or right, we decided to give it a try. One day, we went to work as usual without breakfast and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping.

Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data below, solve our problem.

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7
10	7
7	5
6	6
9	5

Figure 7.38

Figure Descriptions

Figure 7.17: This is a normal distribution curve with mean equal to zero. A vertical line near the tail of the curve to the right of zero extends from the axis to the curve. The region under the curve to the right of the line is shaded representing p-value = 0.0004.

Figure 7.20: This is a normal distribution curve with mean equal to zero. A vertical line near the tail of the curve to the right of zero extends from the axis to the curve. The region under the curve to the right of the line is shaded representing p-value = 0.1207.

References

Figures

Figure 7.17: Figure 10.22 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/10-solutions#fs-idm5499280-solution

Figure 7.20: Figure 10.23 from OpenStax Statistics (2020) (CC BY 4.0). Retrieved from https://openstax.org/books/statistics/pages/10-solutions#fs-idm5499280-solution

Text

Data from *Educational Resources*, December catalog.

Data from Hilton Hotels. Available online at http://www.hilton.com (accessed June 17, 2013).

Data from Hyatt Hotels. Available online at http://hyatt.com (accessed June 17, 2013).

Data from Statistics, United States Department of Health and Human Services.

Data from Whitney Exhibit on loan to San Jose Museum of Art.

Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).

Data from the Chancellor's Office, California Community Colleges, November 1994.

"State of the States," Gallup, 2013. Available online at http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive (accessed June 17, 2013).

"West Nile Virus," Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (accessed June 17, 2013).

Notes

1. Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).
2. Hinduja, Sameer. "Sexting Research and Gender Differences," Cyberbulling Research Center, 2013. Available online at http://cyberbullying.us/blog/sexting-research-and-gender-differences/ (accessed June 17, 2013).
3. "Smart Phone Users, By the Numbers," Visually, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed June 17, 2013).
4. "List of current United States Senators by Age," Wikipedia. Available online at http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age (accessed June 17, 2013).
5. "Texas Crime Rates 1960–1012," FBI, Uniform Crime Reports, 2013. Available online at: http://www.disaster-center.com/crime/txcrime.htm (accessed June 17, 2013).
6. "State of the States," Gallup, 2013. Available online at http://www.gallup.com/poll/125066/StateStates.aspx?ref=interactive (accessed June 17, 2013).
7. Data from Statistics, United States Department of Health and Human Services.
8. Data from Whitney Exhibit on loan to San Jose Museum of Art
9. Data from the Chancellor's Office, California Community Colleges, November 1994.
10. "West Nile Virus," Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (accessed June 17, 2013).
11. Data from Educational Resources, December catalog.
12. "State-Specific Prevalence of Obesity AmongAduls—Unites States, 2007," MMWR, CDC. Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm (accessed June 17, 2013).
13. "Smart Phone Users, By the Numbers," Visually, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed June 17, 2013).

Glossary

***t*-distribution**

A family of distributions, dependent on degrees of freedom, similar to the normal distribution but with more variability built in

Alternative hypothesis

A working hypothesis that is contradictory to the null hypothesis

Anecdotal evidence

Evidence that is based on personal testimony and collected informally

Association

A relationship between variables

Bernoulli trial

An experiment with the following characteristics:

- There are only two possible outcomes (called “success” and “failure”) for each trial.
- The probability (p) of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Bimodal distribution

A distribution that has two modes

Binomial distribution

A random variable that counts the number of successes in a fixed number (n) of independent Bernoulli trials each with probability of a success (p)

Bivariate data

Data consisting of two variables, often in search of an association

Blinding

Not telling participants which treatment they are receiving

Block design study

Grouping individuals based on a variable into "blocks" and then randomizing cases within each block to the treatment groups

Case-control study

A study that compares a group that has a certain characteristic to a group that does not, often a retrospective study for rare conditions

Center

The central tendency or most typical value of a dataset

Central limit theorem (CLT)

If there is a population with mean μ and standard deviation σ, and you take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed.

Class midpoint

Found by adding the lower limit and upper limit, then dividing by two

Class width

The difference in consecutive lower class limits

Cluster sampling

A method of sampling where the population has already sorted itself into groups (clusters), and researchers randomly select a cluster and use every individual in the chosen cluster as the sample

Coefficient of determination

A numerical measure of the percentage or proportion of variation in the dependent variable (y) that can be explained by the independent variable (x)

Cohort study

Longitudinal study where a group of people (typically sharing a common factor) are studied and data is collected for a purpose

Complement

The complement of an event consists of all outcomes in a sample space that are NOT in the event.

Completely randomized study

Dividing participants into treatment groups randomly

Conditional probability

The likelihood that an event will occur given knowledge of another event

Confidence interval

An interval built around a point estimate for an unknown population parameter

Confounding (lurking, conditional) variable

A variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Contingency (two-way) table

A table in a matrix format that displays the frequency distribution of different variables

Continuity correction

When statisticians add or subtract .5 to values to improve approximation

Continuous random variable

A random variable (RV) whose outcomes are measured as an uncountable, infinite number of values

Control group

A group in a randomized experiment that receives no (or inactive) treatment but is otherwise managed exactly as the other groups

Controlled (designed) experiment

Type of experiment where variables are manipulated and data is collected in a controlled setting

Convenience sampling

Selecting individuals that are easily accessible and may result in biased data

Correlation coefficient

A numerical measure that provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y

Critical value

Point on a distribution that acts as a cut-off value for accepting or rejecting the null hypothesis

Cross-sectional study

Data collection on a population at one point in time (often prospective)

Cumulative distribution function (CDF)

A function that gives the probability that a random variable takes a value less than or equal to x

Cumulative relative frequency

The sum of the relative frequencies for all values that are less than or equal to the given value

Data

Actual values (numbers or words) that are collected from the variables of interest

Data analysis process

Process of collecting, organizing, and analyzing data

Degrees of freedom

The number of objects in a sample that are free to vary

Descriptive statistics

Methods of organizing, summarizing, and presenting data

Designed (controlled) experiment

Data collection where variables are manipulated in a controlled setting

Difference in means

The difference in the means of two independent populations

Discrete random variable

A random variable that takes on a countable amount of values

Distribution

The possible values a variable can take on and how often it does so

Double-blind study

The act of blinding both the subjects of an experiment and the researchers who work with the subjects

Empirical rule

Roughly 68% of values are within one standard deviation of the mean, roughly 95% of values are within two standard deviations of the mean, and 99.7% of values are within three standard deviations of the mean

Event

An outcome or subset of outcomes of an experiment in which you are interested

Expected value

Mean of a random variable

Experimental unit

Any individual or object to be measured

Explanatory variable

The independent variable in an experiment; the value controlled by researchers

Extrapolation

The process of predicting outside of the observed *x* values

Factors

Variables in an experiment

Frequency

The number of times a value occurs in the data

Graphical descriptive methods

Organizing, summarizing, or presenting data visually in graphs, figures, or charts

Hypothesis testing

A decision-making procedure for determining whether sample evidence supports a hypothesis

Independent

The occurrence of one event has no effect on the probability of the occurrence of another event.

Individuals

The person, animal, item, place, etc. about which we collect information

Inferential statistics

The facet of statistics dealing with using a sample to generalize (or infer) about the population

Influential points

Observed data points that do not follow the trend of the rest of the data and have a large influence on the calculation of the regression line

Intersection

The shared or common outcomes of two events (i.e., elements are in both A *and* B)

Interval scale level

Quantitative data where the difference or gap between values is meaningful

Law of large numbers

As the number of trials in a probability experiment increases, the relative frequency of an event approaches the theoretical probability

Levels

Certain values of variables in an experiment

Linear regression

A mathematical model of a linear association

Longitudinal study

Collecting data multiple times on the same individuals over a period of time, usually in fixed increments

Lower class limit

The lower end of a bin or class in a frequency table or histogram

Margin of error (MoE)

How much a point estimate can be expected to differ from the true population value; made up of the standard error multiplied by the critical value

Matched pairs design

Very similar individuals (or even the same individual) receive two different treatments (or treatment vs. control), then the results are compared

Mean (average)

A number that measures the central tendency of the data

Measures of location

A measure of an observation's standing relative to the rest of the dataset

Median

The middle number in a sorted list

Modality

How many peaks or clusters there appear to be in a quantitative distribution

Mode

The most frequently occurring value

Mutually exclusive (disjoint)

Two events that cannot happen at the same time; sharing no common outcomes

Nominal scale level

Categorical data where the the categories have no natural, intuitive, or obvious order

Normal (Gaussian) distribution

A commonly used symmetric, unimodal, bell-shaped, and continuous probability distribution

Null hypothesis

The claim that is assumed to be true and is tested in a hypothesis test

Numerical descriptive methods

Numbers that summarize some aspect of a dataset, often calculated

Observational study

Data collection where no variables are manipulated

Ordinal scale level

Categorical data where the the categories have a natural or intuitive order

Outcome

A particular result of an experiment

Outlier

An observation that stands out from the rest of the data significantly

p-value

The probability that an event will occur, assuming the null hypothesis is true

Parameter

A number that is used to represent a population characteristic and can only be calculated as the result of a census

Placebo

An inactive treatment that has no real effect on the explanatory variable

Point estimate

The value that is calculated from a sample used to estimate an unknown population parameter

Point estimation

Using sample data to calculate a single statistic as an estimate of an unknown population parameter

Pooled proportion

Estimate of the common value of p1 and p2

Population

The whole group of individuals who can be studied to answer a research question

Population mean

The arithmetic mean, or average, of a population

Population mean difference

The mean of the differences in a matched pairs design

Population proportion

The number of individuals that have a characteristic of interest divided by the total number in the population

Power

The probability of failing to reject a true hypothesis

Probability

The study of randomness; a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Probability density function (PDF)

A function that defines a continuous random variable and the likelihood of an outcome

Probability experiment

A random experiment where the result is not predetermined

Probability mass function (PMF)

A function that gives the probability that a discrete random variable is exactly equal to some value (x)

Probability model

A mathematical representation of a random process that lists all possible outcomes and assigns probabilities to each

Prospective study

Collecting information as events unfold

Qualitative data

Data that describes qualities or puts individuals into categories; also known as categorical data

Quantile

Points in a distribution that relate to the rank order of values in that distribution

Quantitative continuous data

Data produced by a variable that takes on an uncountable, infinite number of values

Quantitative data

Numerical data with a mathematical context

Quantitative discrete data

Data produced by a variable that takes on a countable number of values

Random variable

A representation of a probability model

Ratio scale level

Quantitative data where the difference or gap between values is meaningful AND has a true 0 value

Relative frequency

The percentage, proportion, or ratio of the frequency of a value of the data to the total number of outcomes

Repeated measures

When an individual goes through a single treatment more than once

Residual (error)

A residual measures the vertical distance between an observation and the predicted point on a regression line

Response variable

The dependent variable in an experiment; the value that is measured for change at the end of an experiment

Retrospective study

Collecting or using data after events have taken place

Robust

Not affected by violations of assumptions such as outliers

Sample

A subset of the population studied

Sample mean

The arithmetic mean, or average, of a dataset

Sample proportion

The number of individuals that have a characteristic of interest divided by the total number in the sample, often found from categorical data

Sample space

The set of all possible outcomes of an experiment

Sampling bias

Bias resulting from all members of the population not being equally likely to be selected

Sampling distribution

The probability distribution of a statistic at a given sample size

Sampling variability

The idea that samples from the same population can yield different results

Shape

The visual appearance of a dataset

Significance level

Probability that a true null hypothesis will be rejected, also known as type I error and denoted by α

Simple random sample (SRS)

Each member of the population is equally likely to be chosen for a sample of a given sample size *and* each sample is equally likely to be chosen

Slope

Tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average

Spread

The level of variability or dispersion of a dataset; also commonly known as variation/variability

Standard deviation

The average distance (deviation) of each observation from the mean

Standard error

The standard deviation of a sampling distribution

Standard normal distribution (SND)

A normal random variable with a mean of 0 and standard deviation of 1 which z-scores follow; denoted N(0, 1)

Statistic

A number calculated from a sample

Statistical inference

Using information from a sample to answer a question, or generalize, about a population

Statistically significant

Finding sufficient evidence that the observed effect is not just due to variability, often from rejecting the null hypothesis

Stratified sampling

Dividing a population into groups (strata) and then using simple random sampling to identify a proportionate number of individuals from each

Systematic (probability) sampling

Using some sort of pattern or probability-based method for choosing your sample

Test statistic

A measure of the difference between observations and the hypothesized (or claimed) value

Treatment combinations (interactions)

Combinations of levels of variables in an experiment

Treatments

Different values or components of the explanatory variable applied in an experiment

Tree diagram

Diagram that helps calculate and organize the number of possible outcomes of an event or problem

Type I error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true

Type II error

Erroneously rejecting a true null hypothesis or erroneously failing to reject a false null hypothesis

Uniform distribution

A probability distribution in which all outcomes are equally likely

Union

The set of all outcomes in two (or more) events (i.e., elements are in A *or* B)

Upper class limit

The upper end of a bin or class in a frequency table or histogram

Values

Possible observations of the variable

Variable

A characteristic of interest for each person or object in a population

Variance

The square of the standard deviation; a computational step along the way to calculating the standard deviation

Variation

The level of variability or dispersion of a dataset; also commonly known as spread or variability

Venn diagram

A diagram that shows all possible relations between a collection of different sets

y-intercept

The value of y when x is 0 in a regression equation

z-score

A measure of location that tells us how many standard deviations a value is above or below the mean

Tables

- **Student *t* table** (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm)
- **Normal table** (https://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm)
- **Chi-Square table** (https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm)
- **F-table** (https://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm)
- **95% Critical Values of the Sample Correlation Coefficient Table** (http://commres.net/wiki/_media/correlationtable.pdf)

Version Notes

This text differs from the existing *Significant Statistics Beta Version* (https://pressbooks.lib.vt.edu/introstatistics). Any changes between these two versions are detailed below.

- The simple linear regression material (old chapter 9) has been moved after univariate descriptives and reframed as bivariate descriptives (new chapter 3)
- Some bivariate quantitative response vs categorical grouping variable graphical methods have been added (new section 3.1)
- Some bivariate categorical graphical methods have been added (new section 3.1) and contingency tables (old section 3.2) have been moved
- Inference for regression has been taken out (old section 9.5)
- Compound events (old section 3.3) have been taken out
- Some of the old chapter 3 (basic probability), old chapter 4 (discrete random variables & binomial), and old chapter 5 (continuous random variables, uniform, normal, and normal approximation) have been combined into the new chapter 4
- Old chapter 6 → New chapter 5
- Old chapter 7 → New chapter 6
- Old chapter 8 → New chapter 7
- Quite a few formatting, wording, and grammatical changes across all chapters
- Professional copyediting
- Peer-reviewed by two reviewers