MGT 235: INTRODUCTORY BUSINESS STATISTICS



Gettysburg College

LibreTexts™

Introductory Business Statistics

Gettysburg College 3rd edition

Original edition published by OpenStax and authored by

Alexander Holmes, The University of Oklahoma
Barbara Illowsky, De Anza College
Susan Dean, De Anza College

Edited using LibreTexts for Gettysburg College Department of Management by

Professors Alice Brawley Newlin and Marta Maras, Gettysburg College

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 05/07/2026

Welcome to MGT 235!

Welcome to MGT 235!

We're glad that you're here. We hope that you will find the knowledge you gain in this class useful not only in your selected major here at Gettysburg, but also beyond the classroom in your day-to-day life. Please know that your instructors are here to support you as you practice these skills and become statistically savvy - keep us informed when you have questions, whether that is in class, in office hours, or by email.

Acknowledgements

Because this book is an open educational resource (OER), we can revise it ourselves as we encounter errors or wish to make updates. We are especially grateful to our students who have helped us catch mistakes and made suggestions for clarifications. These include:

Candice Kong '23
Fadekemi Agboola '24
Brandon Banda '24
Colin Dougherty '24
Maddy Filetti '24
Caleb Monroe '24
Maro Ramadan '24
Regan Rightmire '24
Lauren O'Connell '25
Saheedah Yusuf '25
Beatrice Slevin-Trigo '26

Thank you!

TABLE OF CONTENTS

Welcome to MGT 235!

Licensing

1: Sampling and Data

- 1.1: Introduction
- 1.2: Definitions of Statistics, Probability, and Key Terms
- 1.3: Data, Sampling, and Variation in Data and Sampling
- 1.4: Levels of Measurement
- 1.5: Experimental Design and Ethics
- 1.6: Chapter 1 Key Terms
- 1.7: Chapter 1 Review
- 1.8: Chapter 1 Homework
- 1.9: Chapter 1 Solutions
- 1.10: Chapter 1 References

2: Descriptive Statistics

- 2.1: Introduction
- 2.2: Display Data
- 2.3: Measures of the Location of the Data
- 2.4: Measures of the Center of the Data
- 2.5: Sigma Notation and Calculating the Arithmetic Mean
- 2.6: Skewness and the Mean, Median, and Mode
- 2.7: Measures of the Spread of the Data
- 2.8: Chapter 2 Key Terms
- 2.9: Chapter 2 Review
- 2.10: Chapter 2 Formula Review
- 2.11: Chapter 2 Homework
- 2.12: Chapter 2 Solutions

3: Probability

- 3.1: Introduction to Probability
- 3.2: Probability Terminology
- 3.3: Independent and Mutually Exclusive Events
- 3.4: Two Basic Rules of Probability
- 3.5: Contingency Tables and Probability Trees
- 3.6: Chapter 3 Key Terms
- 3.7: Chapter 3 Review
- 3.8: Chapter 3 Formula Review
- 3.9: Chapter 3 Homework
- 3.10: Chapter 3 Solutions
- 3.11: Chapter 3 References

4: The Normal Distribution

- 4.1: Introduction
- 4.2: The Standard Normal Distribution
- 4.3: Using the Normal Distribution
- 4.4: Chapter 4 Key Terms
- 4.5: Chapter 4 Review
- 4.6: Chapter 4 Formula Review
- 4.7: Chapter 4 Homework
- 4.8: Chapter 4 Solutions
- 4.9: Chapter 4 References

5: The Central Limit Theorem

- 5.1: Introduction to the Central Limit Theorem
- 5.2: The Central Limit Theorem for Sample Means
- 5.3: Using the Central Limit Theorem
- 5.4: Chapter 5 Key Terms
- 5.5: Chapter 5 Review
- 5.6: Chapter 5 Formula Review
- 5.7: Chapter 5 Homework
- 5.8: Chapter 5 Solutions
- 5.9: Chapter 5 References

6: Confidence Intervals

- 6.1: Introduction
- 6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size
- 6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case
- 6.4: A Confidence Interval for A Population Proportion
- 6.5: Chapter 6 Key Terms
- 6.6: Chapter 6 Review
- 6.7: Chapter 6 Formula Review
- 6.8: Chapter 6 Homework
- 6.9: Chapter 6 Solutions
- 6.10: Chapter 6 References

7: Hypothesis Testing with One Sample

- 7.1: Introduction to Hypothesis Testing
- 7.2: Null and Alternative Hypotheses
- 7.3: Outcomes and Type I and Type II Errors
- 7.4: Distribution Needed for Hypothesis Testing
- 7.5: Full Hypothesis Test Examples
- 7.6: Chapter 7 Key Terms
- 7.7: Chapter 7 Review
- 7.8: Chapter 7 Formula Review
- 7.9: Chapter 7 Homework
- 7.10: Chapter 7 Solutions
- 7.11: Chapter 7 References

8: Hypothesis Testing with Two Samples

- 8.1: Introduction
- 8.2: Comparing Two Independent Population Means
- 8.3: Cohen's Standards for Small, Medium, and Large Effect Sizes
- 8.4: Comparing Two Independent Population Proportions
- 8.5: Matched or Paired Samples
- 8.6: Chapter 8 Key Terms
- 8.7: Chapter 8 Review
- 8.8: Chapter 8 Formula Review
- 8.9: Chapter 8 Homework
- 8.10: Chapter 8 Solutions
- 8.11: Chapter 8 References

9: F-Distribution and One-Way ANOVA

- 9.1: Introduction
- 9.2: One-Way ANOVA
- 9.3: The F-Distribution and the F-Ratio
- 9.4: Chapter 9 Key Terms
- 9.5: Chapter 9 Review
- 9.6: Chapter 9 Formula Review
- 9.7: Chapter 9 Homework
- 9.8: Chapter 9 Solutions
- 9.9: Chapter 9 References

10: Linear Regression and Correlation

- 10.1: Introduction
- 10.2: The Correlation Coefficient r
- 10.3: Testing the Significance of the Correlation Coefficient
- 10.4: Linear Equations
- 10.5: The Regression Equation
- 10.6: How to Use Microsoft Excel® for Regression Analysis
- 10.7: Chapter 10 Key Terms
- 10.8: Chapter 10 Review
- 10.9: Chapter 10 Homework
- 10.10: Chapter 10 Solutions

11: Appendices

- 11.1: A - Statistical Tables
- 11.2: B - Mathematical Phrases, Symbols, and Formulas
- 11.3: C - Reporting Statistics in APA Style

Index

Glossary

Detailed Licensing

Licensing

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

CHAPTER OVERVIEW

1: Sampling and Data

1.1: Introduction
1.2: Definitions of Statistics, Probability, and Key Terms
1.3: Data, Sampling, and Variation in Data and Sampling
1.4: Levels of Measurement
1.5: Experimental Design and Ethics
1.6: Chapter 1 Key Terms
1.7: Chapter 1 Review
1.8: Chapter 1 Homework
1.9: Chapter 1 Solutions
1.10: Chapter 1 References

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

1.1: Introduction



Figure 1.1.1 We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

You are probably asking yourself the question, "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 crime, sports, education, politics, 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. Think about your chosen profession. 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 words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

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

1.2: Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives. **Data** represent all the pieces of information or observations collected on characteristics of our interest (actual values of the variable). They may be numbers or they may be words. **Datum** is a single value. Data can come from a population or a sample.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. 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.

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, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 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 whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a successful surgery or medical treatment. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations 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 persons, things, or objects 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. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. 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 poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, 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, in this case the mean. 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 per student over all the 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 the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, or random variable, usually denoted by capital letters such as X and Y, is a characteristic or measurement that can be determined for each member of a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. 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 some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note

The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example 1.2.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Answer

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term: the population mean.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Exercise 1.2.1

Determine what the key terms refer to in 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, respectively.

Example 1.2.2

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population ____ 2. Statistic ____ 3. Parameter ____ 4. Sample ____ 5. Variable ____ 6. Data ____

a. all students who attended the college last year
b. the cumulative GPA of one student who graduated from the college last year
c. 3.65, 2.80, 1.50, 3.90
d. a group of students who graduated from the college last year, randomly selected
e. the average cumulative GPA of students who graduated from the college last year
f. all students who graduated from the college last year
g. the average cumulative GPA of students in the study who graduated from the college last year

Answer

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

Example 1.2.3

Determine what the key terms refer to in the following study.

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 an automobile crash on test dummies. Here is the criterion they used. 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.

Answer

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Example 1.2.4

Determine what the key terms refer to in the following study.

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 the number in the sample who have been involved in a malpractice lawsuit.

Answer

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

This page titled 1.2: Definitions of Statistics, Probability, and Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by .

1.3: Data, Sampling, and Variation in Data and Sampling

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

- Qualitative (categorical)
- Quantitative (numerical)

Qualitative data are the result of categorizing or describing attributes of a population. **Qualitative data** are also often called **categorical** data. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative (categorical) data. Qualitative (categorical) data are generally described by 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+. Researchers often prefer to use quantitative (numerical) data over qualitative (categorical) data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Note that it's possible to collect data as numbers and then report it categorically. For example, suppose the exact quiz scores for each student are recorded throughout the term (e.g., 97%). At the end of the term, the quiz scores are reported as A, B, C, D, or F, which are categories.

Quantitative (numerical) data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

Example 1.3.1

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. The numbers of books (three, four, two, and one) are the quantitative discrete data.

Exercise 1.3.1

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?

Example 1.3.2

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

Exercise 1.3.2

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?

Example 1.3.3

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

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

Answer

One Possible Solution:

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

Try to identify additional data sets in this example.

Example 1.3.4

The data are the colors of backpacks. Again, you sample the same 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. The colors red, black, black, green, and gray are qualitative (categorical) data.

Exercise 1.3.3

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?

Example 1.3.5

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. the distance from your home to the nearest grocery store
d. the number of classes you take per school year
e. the type of calculator you use
f. weights of sumo wrestlers
g. number of correct answers on a quiz
h. IQ scores (This may cause some discussion.)

Answer

Items a, d, and g are quantitative discrete; items c, f, and h are quantitative continuous; items b and e are qualitative, or categorical.

Exercise 1.3.4

Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

Example 1.3.6

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the bar chart in Figure 1.3.1. What type of data does this graph show?

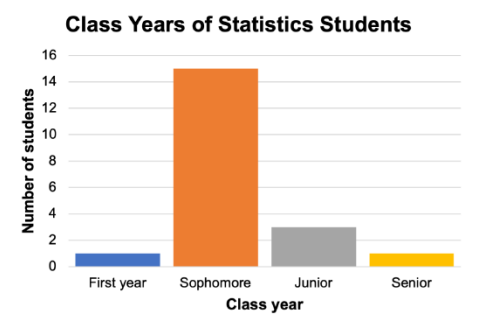


Figure 1.3.1

Answer

This bar chart shows the students in each year, which is **qualitative (or categorical) data**.

Exercise 1.3.5

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.

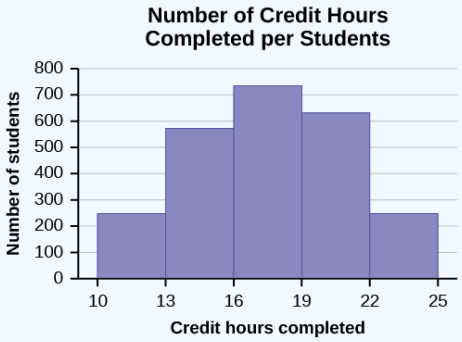


Figure 1.3.2

What type of data does this graph show?

Qualitative Data Discussion

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 comparing the same categories in the colleges easier. 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.

Table 1.3.1: Spring Term 2010 (Census day)

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%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. A **bar graph**, where the length of the bar for each category is proportional to the number or percent of individuals in each category, is the appropriate choice to display data from categorical variables. Bars may be vertical or horizontal, and as seen in Figure 1.3.4 below, **clustered bar charts** can be created to show group comparisons.

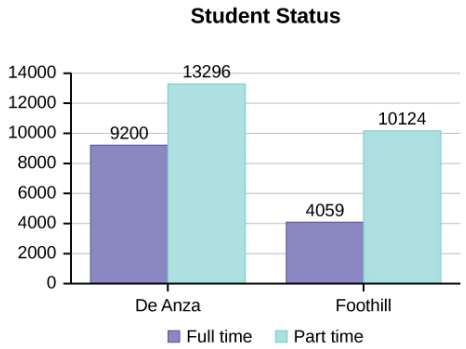


Figure 1.3.3

Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the table below, the percentages add to more than 100% because students can be in more than one category. A bar graph can still be used to compare the relative size of the categories.

Table 1.3.2: De Anza College Spring 2010

Characteristic/category	Percent
Full-time students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

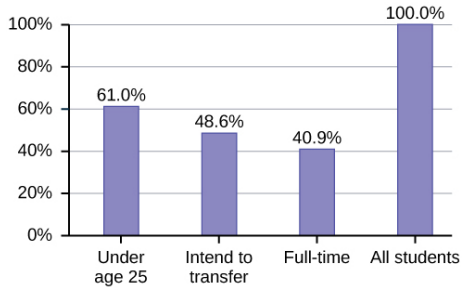


Figure 1.3.4

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. As above, though, a bar chart can still be created to display this data.

Table 1.3.3: Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

	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%

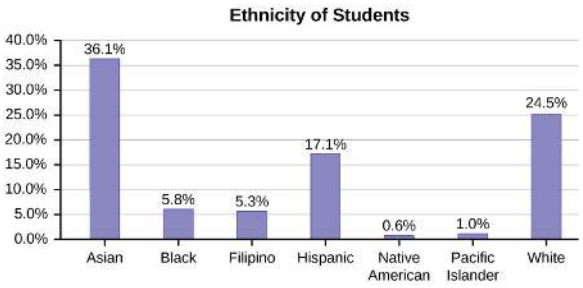


Figure 1.3.5

The following graph 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 (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure 1.3.7 is a Pareto chart. The **Pareto chart** has the bars sorted from largest to smallest and is sometimes easier to read and interpret.

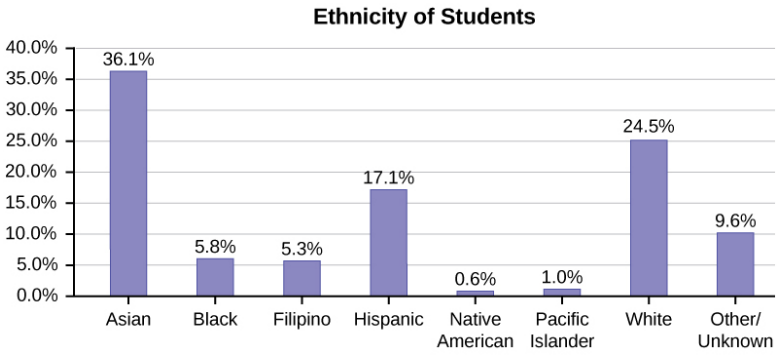


Figure 1.3.6: Bar Graph with Other/Unknown Category

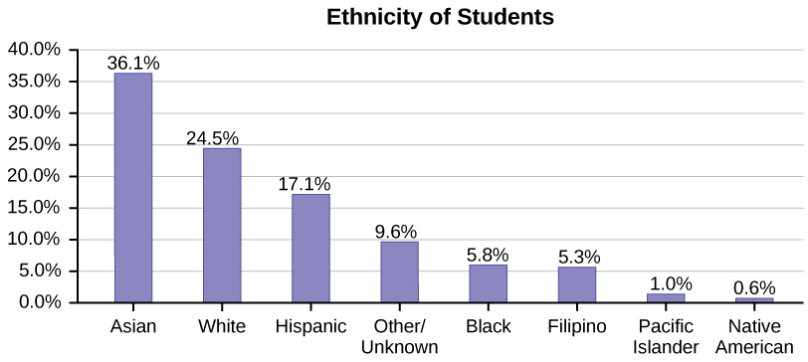


Figure 1.3.7: Pareto Chart With Bars Sorted by Size

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. A sample should have the same characteristics as the population it is representing. Most 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 of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. 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.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.

To choose a **stratified sample**, divide the population into groups called strata and then (usually) 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 stratum (each department) to get a stratified random sample. For the stratified sample to be proportionate, proportions of the groups in the sample need to be equal to their proportions in the population. If a specific department accounts for 10% of the college population, it should have the same (10%) representation in the sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

There are also situations where it is reasonable to collect a **disproportional stratified sample**. In these cases, you still divide the population into groups called strata, but set your proportions according to a desire to "over"-sample a particular strata. This is most useful in situations where one strata comprises a very small percentage of the total population.

For example, let's say the US National Park Service (NPS) wants to compare local and nationwide public opinion about Yellowstone National Park, and they plan to do this by comparing responses from Wyoming residents (local to Yellowstone) to responses from other states. The US state of Wyoming has a population of approximately 600,000 residents (as of 2020). This state represents a very small percentage (0.2%) of the total US population of about 328,200,000.

With that in mind, suppose the NPS were sampling 1,000 people total. If they wanted to conclude something meaningful and trustworthy about each group - Wyoming compared to the rest of the US - then a proportional stratified sample would be useless. It would include only 2 people (sample size of 1,000 x the 0.2% of the population living in Wyoming = 2 respondents)! Instead, it would be up to the research team to determine what proportions should be used for each strata. Perhaps they would decide on setting each strata's proportion at 50%. Note that this would still be called "disproportionate" stratified sampling because the strata proportions do not match the population proportions.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the

clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every k^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then calculate the skip number **k** by dividing the number of individuals in the population by the number of individuals needed in the sample. Choose every fiftieth name thereafter starting with the one that was randomly selected until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a rather simple sampling method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** 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 (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

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, in most populations, simple random sampling is done without replacement. Surveys are typically done without replacement. That is, 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 for this sample 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 1000 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 999/10,000 and 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 ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

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

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and **nonsampling 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 nonsampling errors. A defective counting device can cause a nonsampling 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 a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Critical Evaluation

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

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: Collecting data or asking questions in a way that influences the response
- Non-response or refusal of 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 relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: Improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Example 1.3.7

A study is done to determine the average tuition that San Jose State 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?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Answer

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Example 1.3.8

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Answer

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example 1.3.9

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 and taking every fifth senior citizen on the list, for 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.

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Answer

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Answer

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

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

c. Is the sample biased?

Answer

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Exercise 1.3.6

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 music 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 is characteristic of) the entire 20,000 listener population?

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 (in ounces) of beverage:

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 make 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.

Be aware that as you take data, your data 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 reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that 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, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples 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. But still, their samples would be, in all likelihood, different from each other. This variability in samples cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations, usually given the symbol n) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. Later we will find that even much smaller sample sizes will give very good results. You will learn why when you study confidence intervals.

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

This page titled 1.3: Data, Sampling, and Variation in Data and Sampling is shared under a CC BY license and was authored, remixed, and/or curated by .

1.4: Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as needed for your research and presentation purposes.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend 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. They are (from lowest to highest level):

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

Data that is measured using a **nominal scale** is **qualitative (categorical)**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. The data are the names of 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. The 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 national parks in the United States 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 measured. Like the nominal scale data, ordinal scale data cannot be used in most calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. 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 done. 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.

Frequency

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; 3.

Table 1.4.1 lists the different data values in ascending order and their frequencies.

Data value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

Table 1.4.1 Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to Table 1.4.1, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

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.

Data value	Frequency	Relative frequency
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Table 1.4.2 Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of Table 1.4.2 is $\frac{20}{20}$, or 1.

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 Table 1.4.3.

Data value	Frequency	Relative frequency	Cumulative relative frequency
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Table 1.4.3 Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Note

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

Table 1.4.4 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Heights (inches)	Frequency	Relative frequency	Cumulative relative frequency
59.96–61.95	5	$\frac{5}{10} = 0.05$	0.05
61.96–63.95	3	$\frac{3}{100} = 0.03$	0.05 + 0.03 = 0.08
63.96–65.95	15	$\frac{15}{100} = 0.15$	0.08 + 0.15 = 0.23
65.96–67.95	40	$\frac{40}{100} = 0.40$	0.23 + 0.40 = 0.63
67.96–69.95	17	$\frac{17}{100} = 0.17$	0.63 + 0.17 = 0.80
69.96–71.95	12	$\frac{12}{100} = 0.12$	0.80 + 0.12 = 0.92
71.96–73.95	7	$\frac{7}{100} = 0.07$	0.92 + 0.07 = 0.99
73.96–75.95	1	$\frac{1}{100} = 0.01$	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Table 1.4.4 Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.96 to 61.95 inches
- 61.96 to 63.95 inches
- 63.96 to 65.95 inches
- 65.96 to 67.95 inches
- 67.96 to 69.95 inches
- 69.96 to 71.95 inches
- 71.96 to 73.95 inches
- 73.96 to 75.95 inches

In this sample, there are **five** players whose heights fall within the interval 59.96–61.95 inches, **three** players whose heights fall within the interval 61.96–63.95 inches, **15** players whose heights fall within the interval 63.96–65.95 inches, **40** players whose heights fall within the interval 65.96–67.95 inches, **17** players whose heights fall within the interval 67.96–69.95 inches, **12** players whose heights fall within the interval 69.96–71.95, **seven** players whose heights fall within the interval 71.96–73.95, and **one** player whose heights fall within the interval 73.96–75.95.

Example 1.4.1

From Table 1.4.4, find the percentage of heights that are up to 65.95 inches.

Exercise 1.4.1

Table 1.4.5 shows the amount, in inches, of annual rainfall in a sample of towns.

Table 1.4.5

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
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	

From Table 1.4.5, find the percentage of rainfall that is less than 9.01 inches.

Example 1.4.2

From Table 1.4.4, find the percentage of heights that fall between 61.96 and 65.95 inches.

Answer

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

Exercise 1.4.2

From Table 1.4.5, find the percentage of rainfall that is between 6.99 and 13.05 inches.

Example 1.4.3

Use the heights of the 100 male semiprofessional soccer players in Table 1.4.4. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.96 to 71.95 inches is: ____.
2. The percentage of heights that are from 67.96 to 73.95 inches is: ____.
3. The percentage of heights that are more than 65.95 inches is: ____.
4. The number of players in the sample who are between 61.96 and 71.95 inches tall is: ____.
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

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.

Answer

1. 29%
2. 36%
3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Example 1.4.4

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. Table 1.4.6 was produced:

Table 1.4.6 Frequency of Commuting Distances

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	$\frac{4}{19}$	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

1. Is the table correct? If it is not correct, what is wrong?
2. 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.
3. What fraction of the people surveyed commute five or seven miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Answer

1. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
3. $\frac{5}{19}$
4. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Exercise 1.4.3

Table 1.4.5 represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Example 1.4.5

Table 1.4.7 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
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Table 1.4.7

Answer the following questions.

1. What is the frequency of deaths measured from 2006 through 2009?
2. What percentage of deaths occurred after 2009?
3. What is the relative frequency of deaths that occurred in 2003 or earlier?
4. What is the percentage of deaths that occurred in 2004?
5. What kind of data are the numbers of deaths?
6. 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?

Answer

1. 97,118 (11.8%)
2. 41.6%
3. 67,092/823,356 or 0.081 or 8.1 %
4. 27.8%
5. Quantitative discrete
6. Quantitative continuous

Exercise 1.4.4

Table 1.4.8 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		

Table 1.4.8

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.

This page titled 1.4: Levels of Measurement is shared under a CC BY license and was authored, remixed, and/or curated by .

1.5: Experimental Design and Ethics

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? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

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 **independent variable** or **explanatory variable**. The affected variable is called the **dependent variable** or **response variable**: stimulus, response. 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 are called **treatments**. An **experimental unit** is a single object or individual to be measured.

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

Additional variables that can cloud a study are called **lurking variables**. In order to find that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to 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. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can test a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation 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.* (McClung & Collins, 2013, p. 382)

When participation in a study prompts a physical response from a participant, it is 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 which 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 blinded.

Example 1.5.1

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

a. Describe the explanatory and response variables in this study.
b. What are the treatments?
c. Identify any lurking variables that could interfere with this study.

d. Is it possible to use blinding in this study?

Answer

The explanatory variable is scent, and the response variable is the time it takes to complete the maze. There are two treatments: a floral-scented mask and an unscented mask. 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. 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, though. The researcher who is observing a subject will not know which mask is being worn.

This page titled 1.5: Experimental Design and Ethics is shared under a CC BY license and was authored, remixed, and/or curated by .

1.6: Chapter 1 Key Terms

Average

also called mean or arithmetic mean; a number that describes the central tendency of the data

Blinding

not telling participants which treatment a subject is receiving

Categorical Variable

variables that take on values that are names or labels

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Control Group

a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Discrete Random Variable

a random variable (RV) whose outcomes are counted

Double-blinding

the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Experimental Unit

any individual or object to be measured

Explanatory Variable

the **independent variable** in an experiment; the value controlled by researchers

Frequency

the number of times a value of the data occurs

Lurking Variable

a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Mathematical Models

a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Numerical Variable

variables that take on values that are indicated by numbers

Observational Study

a study in which the independent variable is not manipulated by the researcher

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Placebo

an inactive treatment that has no real effect on the explanatory variable

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of observations in the category or quantitative interval of interest divided by the total number of observations in the sample

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Assignment

the act of organizing experimental units into treatment groups using random methods

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes

Representative Sample

a subset of the population that has the same characteristics as the population

Response Variable

the **dependent variable** in an experiment; the value that is measured for change at the end of an experiment

Sample

a subset of the population studied

Sampling Bias

not all members of the population are equally likely to be selected

Sampling Error

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of numbers. These randomly selected numbers identify the members of your sample.

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Statistical Models

a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Survey

a study in which data is collected as reported by individuals.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Treatments

different values or components of the explanatory variable applied in an experiment

Variable

a characteristic of interest for each person or object in a population

---

This page titled 1.6: Chapter 1 Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by .

1.7: Chapter 1 Review

1.1 Definitions of Statistics, Probability, and Key Terms

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

1.2 Data, Sampling, and Variation in Data and Sampling

Data are individual items of information that come from a population or sample. Data may be classified as qualitative (categorical), quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

1.3 Levels of Measurement

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no meaningful zero or starting point; the differences can be measured, but ratios cannot be meaningfully calculated and interpreted.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

1.4 Experimental Design

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

This page titled 1.7: Chapter 1 Review is shared under a CC BY license and was authored, remixed, and/or curated by .

1.8: Chapter 1 Homework

1.2 Definitions of Statistics, Probability, and Key Terms

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

1. A fitness center is interested in the mean amount of time a client exercises in the center each week.

2. 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.

3. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

4. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

5. A politician is interested in the proportion of voters in his district who think he is doing a good job.

6. A marriage counselor is interested in the proportion of clients she counsels who stay married.

7. Political pollsters may be interested in the proportion of people who will vote for a particular cause.

8. A marketing company is interested in the proportion of people who will buy a particular product.

Use the following information to answer the next three exercises: 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.

9. What is the population she is interested in?

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

10. 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.

11. 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.

1.3 Data, Sampling, and Variation in Data and Sampling

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), identify the scale of measurement (NOIR) and give an example of the data.

12. number of tickets sold to a concert

13. percent of body fat

14. favorite baseball team

15. time in line to buy groceries

16. number of students enrolled at Evergreen Valley College

17. most-watched television show

18. brand of toothpaste

19. distance to the closest movie theater

20. age of executives in Fortune 500 companies

21. number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

22. "Number of times per week" is what type of data?

a. qualitative (categorical)
b. quantitative discrete
c. quantitative continuous

23. "Duration (amount of time)" is what type of data?

a. qualitative (categorical)
b. quantitative discrete
c. quantitative continuous

24. 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. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result 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.

25. 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. Make the description detailed.

26. 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. Make the description detailed.

27. List some practical difficulties involved in getting accurate results from a telephone survey.

28. List some practical difficulties involved in getting accurate results from a mailed survey.

29. Brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

30. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

31. 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. The sampling method was:

a. simple random
b. systematic
c. stratified

d. cluster

32. Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all 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 rows two and five and then calls on all students in row two and all students in 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 to fill out asking for information about age, as well as about 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 the reaction of voters 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 whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

33. 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. Also, 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 U.S. 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? Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
c. 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.

34. The Well-Being Index is a survey that follows trends of U.S. 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(categorical), quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
c. In the last seven days, on how many days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

35. 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.

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 nonsampling 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. Quota sampling is an example of which sampling method described in this module?

36. 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.

Which of the potential problems with samples discussed in this chapter could explain this connection?

37. 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?”

As of April 25, 11 people responded to this question. Each participant answered “NO!”

Which of the potential problems with samples discussed in this module could explain this connection?

38. 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.”

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.”

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.

1.4 Levels of Measurement

39. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of courses	Frequency	Relative frequency	Cumulative relative frequency
1	30	0.6	
2	15		
3			

Table 1.8.1 Part-time Student Course Loads

a. Fill in the blanks in Table 1.8.1.
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

40. Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in Table 1.8.2.

# 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	

Table 1.8.2 Flossing Frequency for Adults with Gum Disease

a. Fill in the blanks in Table 1.8.2.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

41. Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2; 5; 7; 2; 2; 10; 20; 15; 0; 7; 0; 20; 5; 12; 15; 12; 4; 5; 10 .

Table 1.8.3 was produced.

Table 1.8.3 Frequency of Immigrant Survey Responses

Data	Frequency	Relative frequency	Cumulative relative frequency
0	2	.1053	0.1053
2	3	.1579	0.2632
4	1	.0526	0.3158
5	3	.1579	0.4737
7	2	.1053	0.5789
10	2	.1053	0.6842
12	2	.1053	0.7895
15	1	.0526	0.8421
20	1	.0526	1.0000

a. Fix the errors in Table 1.8.3. Also, 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 U.S. for 5 years.”
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

42. How much time does it take to travel to work? Table 1.8.4 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.

Table 1.8.4

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

43. *Forbes* magazine published data on the best small firms in 2012. These were firms which 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. Table 1.8.5 shows the ages of the chief executive officers for the first 60 ranked firms.

Table 1.8.5

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		

a. What is the frequency for CEO ages between 55 and 64?
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?

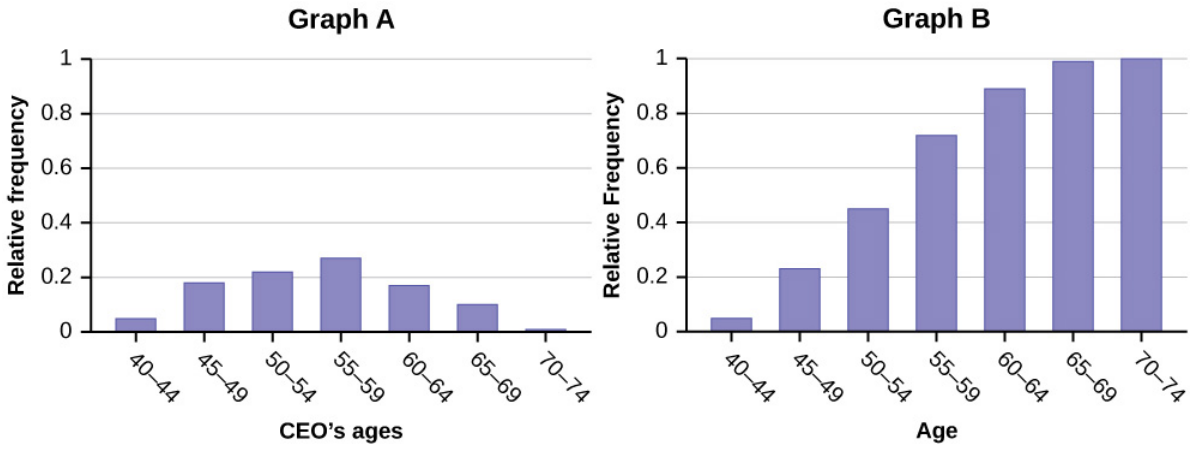


Figure 1.8.1

Use the following information to answer the next two exercises: Table 1.8.6 contains data on hurricanes that have made direct hits on the U.S. 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		

Table 1.8.6 Frequency of Hurricane Direct Hits

44. What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768

b. 0.0659
c. 0.2601
d. Not enough information to calculate

45. What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

1.8: Chapter 1 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.9: Chapter 1 Solutions

2.

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

4.

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

6.

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. X = the number of couples who stay married
f. yes, no

8.

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

10. a

12. quantitative discrete, ratio, 150

14. qualitative, nominal, Oakland A's

16. quantitative discrete, ratio, 11,234 students

18. qualitative, nominal, Crest

20. quantitative continuous, ratio, 47.3 years

22. b

24.

a. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

26. Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

28. Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

30. b

32. convenience; cluster; stratified; systematic; simple random

34.

a. qualitative (categorical)
b. quantitative discrete
c. quantitative discrete
d. qualitative (categorical)

36.

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.

38.

a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

40.

Table 1.19

a.

# flossing per week	Frequency	Relative frequency	Cumulative relative frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

b. 5.00%
c. 93.33%

42. The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

44. b

1.9: Chapter 1 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.10: Chapter 1 References

1.2 Definitions of Statistics, Probability, and Key Terms

The Data and Story Library, lib.stat.cmu.edu/DASL/Stories...stDummies.html (accessed May 1, 2013).

1.3 Data, Sampling, and Variation in Data and Sampling

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).

Data from www.bookofodds.com/Relationsh...-the-President

Dominic Lusinchi, "'President' Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).

"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/ga...9362004.aspx#4 (accessed May 1, 2013).

The Data and Story Library, 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/f...hts.html#focus (accessed May 1, 2013).

Data from San Jose Mercury News

1.4 Levels of Measurement

"State & County QuickFacts," U.S. Census Bureau. quickfacts.census.gov/qfd/download_data.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," infinity.cos.edu/faculty/wood...ata_Levels.htm (accessed May 1, 2013).

Courtney Taylor, "Levels of Measurement," about.com, http://statistics.about.com/od/Helpa...easurement.htm (accessed May 1, 2013).

David Lane. "Levels of Measurement," Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

1.5 Experimental Design and Ethics

"Vitamin E and Health," Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritio...rce/vitamin-e/ (accessed May 1, 2013).

Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, 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-as...s-study-300443 (accessed May 1, 2013).

The Data and Story Library, lib.stat.cmu.edu/DASL/Stories...dLearning.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/earthquak...archives/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).

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...onsumer-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).

This page titled 1.10: Chapter 1 References is shared under a CC BY license and was authored, remixed, and/or curated by .

- **1.9: References** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.
- **Current page** is licensed CC BY 4.0.

CHAPTER OVERVIEW

2: Descriptive Statistics

2.1: Introduction
2.2: Display Data
2.3: Measures of the Location of the Data
2.4: Measures of the Center of the Data
2.5: Sigma Notation and Calculating the Arithmetic Mean
2.6: Skewness and the Mean, Median, and Mode
2.7: Measures of the Spread of the Data
2.8: Chapter 2 Key Terms
2.9: Chapter 2 Review
2.10: Chapter 2 Formula Review
2.11: Chapter 2 Homework
2.12: Chapter 2 Solutions

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

2.1: Introduction



Figure 2.1.1 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. (credit: William Greeson)

Once you have collected data, what will 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. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution 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 data clusters and where there are only a few data values. Newspapers and the Internet 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. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

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

2.2: Display Data

Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. 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 stem.

Example 2.2.1

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

Table 2.2.1 Stem-and-Leaf Graph

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

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% were in the 90s or 100, a fairly high number of As.

Exercise 2.2.1

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.

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it 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 explain outliers, so we will cover them in more detail later.

Example 2.2.2

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

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

Do the data seem to have any concentration of values?

NOTE

The leaves are to the right of the decimal.

Answer

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Table 2.2.2

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Exercise 2.2.2

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

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

Example 2.2.3

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Table 2.2.4 and Table 2.2.5 show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Answer

Table 2.2.3

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
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

Table 2.2.4 Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51	Trump	70

Table 2.2.5 Presidential Age at Death

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		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example 2.2.4, 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.

Example 2.2.4

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table 2.2.6 and in Figure 2.2.1.

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

Table 2.2.6

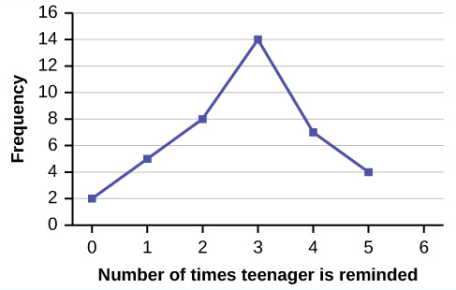


Figure 2.2.1

Exercise 2.2.3

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table 2.2.7. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Table 2.2.7

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in Example 2.2.5 has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 2.2.5

By the end of 2011, Facebook had over 146 million users in the United States. Table 2.2.8 shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

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%

Table 2.2.8

Answer

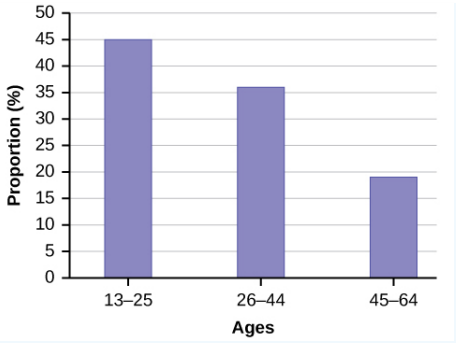


Figure 2.2.2

Exercise 2.2.4

The population in Park City is made up of children, working-age adults, and retirees. Table 2.2.9 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%

Table 2.2.9

Example 2.2.6

The columns in Table 2.2.10 contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, 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%

Table 2.2.10

Answer

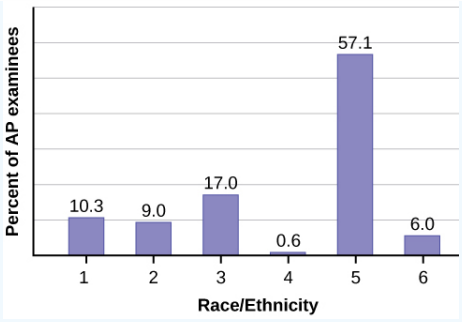


Figure 2.2.3

Exercise 2.2.5

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

Table 2.2.11

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%

Example 2.2.7

Below is a two-way table showing the types of pets owned by men and women:

Table 2.2.12

	Dogs	**Cats**	**Fish**	**Total**
Men	4	2	2	8
Women	4	6	2	12
Total	8	8	4	20

Given these data, calculate the conditional distributions for the subpopulation of men who own each pet type.

Answer

- Men who own dogs = 4/8 = 0.5
- Men who own cats = 2/8 = 0.25
- Men who own fish = 2/8 = 0.25

Note: The sum of all of the conditional distributions must equal one. In this case, 0.5 + 0.25 + 0.25 = 1; therefore, the solution "checks".

Histograms, Frequency Polygons, and Time Series Graphs

For most of the work you do in this book, you will use a histogram to display the quantitative (numerical) data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

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 with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (Remember, frequency is defined as the number of times an answer occurs.) If:

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

then:

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

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

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. 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). We say that 6.05 has more precision. 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 all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). Also, 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.

Example 2.2.8

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; 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 the data with the most decimal places has one decimal (for instance, 61.5), 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.

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. A guideline that is followed by some for the width of a bar or class interval is 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.

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.

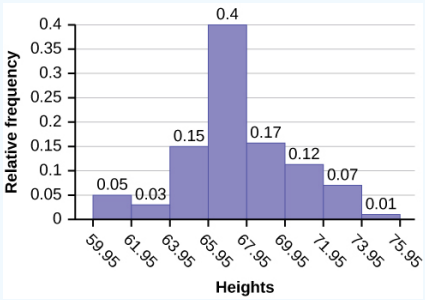


Figure 2.2.4

Exercise 2.2.6

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose 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; 11.5; 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

Example 2.2.9

Create a histogram for the following data: 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; 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. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____ .

Solution

Calculate the number of bars as follows:

$$\frac{6.5 - 0.5}{\text{number of bars}} = 1$$

where 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.

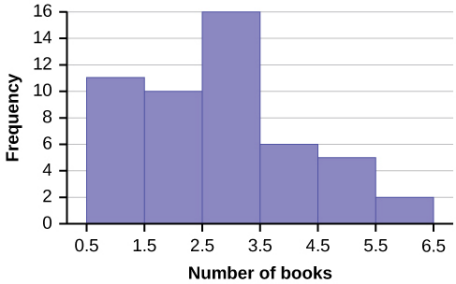


Figure 2.2.5

Example 2.2.10

Using this data set, construct a histogram.

Table 2.2.13

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

Answer



Figure 2.2.6

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. 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.

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

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.

Example 2.2.11

A frequency polygon was constructed from the frequency table below.

Table 2.2.14: Frequency distribution for calculus final test scores

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

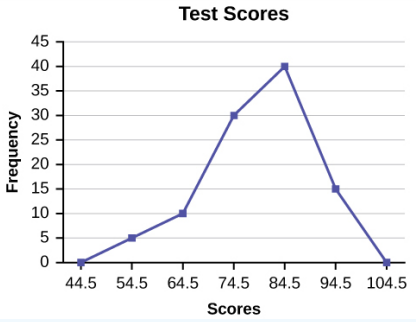


Figure 2.2.7

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.

Exercise 2.2.7

Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in Table 2.2.15.

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	2

Table 2.2.15

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example 2.2.12

We will construct an overlay frequency polygon comparing the scores from Example 2.2.11 with the students' final numeric grade.

Table 2.2.16: Frequency distribution for calculus final test scores

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

Table 2.2.17: Frequency distribution for calculus final grades

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

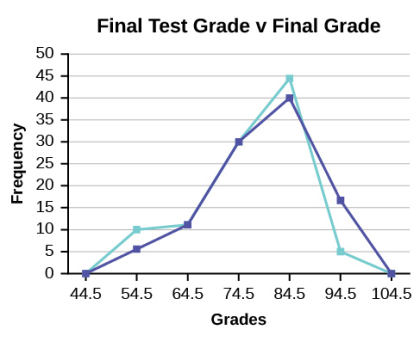


Figure 2.2.8

Constructing a Time Series Graph

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with these 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 reach 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 that of 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 is called a time series graph.

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or 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 2.2.13

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.

Table 2.2.18

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
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	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Table 2.2.19

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
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

Answer

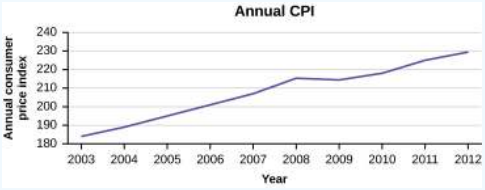


Figure 2.2.9

Exercise 2.2.8

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

Table 2.2.20: CO_2 emissions

Year	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

Uses of 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.

How NOT to Lie with Statistics

It is important to remember that the very reason we develop a variety of methods to present data is to develop insights into the subject of what the observations represent. We want to get a "sense" of the data. Are the observations all very much alike or are they spread across a wide range of values, are they bunched at one end of the spectrum or are they distributed evenly and so on. We are trying to get a visual picture of the numerical data. Shortly we will develop formal mathematical measures of the data, but our visual graphical presentation can say much. It can, unfortunately, also say much that is distracting, confusing and simply wrong in terms of the impression the visual leaves. Many years ago Darrell Huff wrote the book *How to Lie with Statistics*. It has been through 25 plus printings and sold more than one and one-half million copies. His perspective was a harsh one and used many actual examples that were designed to mislead. He wanted to make people aware of such deception, but perhaps more importantly to educate so that others do not make the same errors inadvertently.

Again, the goal is to enlighten with visuals that tell the story of the data. Pie charts have a number of common problems when used to convey the message of the data. Too many pieces of the pie overwhelm the reader. More than perhaps five or six categories ought to give an idea of the relative importance of each piece. This is after all the goal of a pie chart, what subset matters most relative to the others. If there are more components than this then perhaps an alternative approach would be better or perhaps some can be consolidated into an "other" category. Pie charts cannot show changes over time, although we see this attempted all too often. In federal, state, and city finance documents pie charts are often presented to show the components of revenue available to the governing body for appropriation: income tax, sales tax motor vehicle taxes and so on. In and of itself this is interesting information and can be nicely done with a pie chart. The error occurs when two years are set side-by-side. Because the total revenues change year to year, but the size of the pie is fixed, no real information is provided and the relative size of each piece of the pie cannot be meaningfully compared.

Histograms can be very helpful in understanding the data. Properly presented, they can be a quick visual way to present probabilities of different categories by the simple visual of comparing relative areas in each category. Here the error, purposeful or not, is to vary the width of the categories. This of course makes comparison to the other categories impossible. It does embellish the importance of the category with the expanded width because it has a greater area, inappropriately, and thus visually "says" that that category has a higher probability of occurrence.

Time series graphs perhaps are the most abused. A plot of some variable across time should never be presented on axes that change part way across the page either in the vertical or horizontal dimension. Perhaps the time frame is changed from years to months. Perhaps this is to save space or because monthly data was not available for early years. In either case this confounds the presentation and destroys any value of the graph. If this is not done to purposefully confuse the reader, then it certainly is either lazy or sloppy work.

Changing the units of measurement of the axis can smooth out a drop or accentuate one. If you want to show large changes, then measure the variable in small units, penny rather than thousands of dollars. And of course to continue the fraud, be sure that the axis does not begin at zero, zero. If it begins at zero, zero, then it becomes apparent that the axis has been manipulated.

Perhaps you have a client that is concerned with the volatility of the portfolio you manage. An easy way to present the data is to use long time periods on the time series graph. Use months or better, quarters rather than daily or weekly data. If that doesn't get the volatility down then spread the time axis relative to the rate of return or portfolio valuation axis. If you want to show "quick" dramatic growth, then shrink the time axis. Any positive growth will show visually "high" growth rates. Do note that if the growth is negative then this trick will show the portfolio is collapsing at a dramatic rate.

Again, the goal of descriptive statistics is to convey meaningful visuals that tell the story of the data. Purposeful manipulation is fraud and unethical at the worst, but even at its best, making these type of errors will lead to confusion on the part of the analysis.

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

2.3: Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median, Mdn, is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as 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.

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

The **median** is a number that measures the "center" 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, add the two values together and divide by two.

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

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

Quartiles are numbers that separate the data into quarters. Quartiles 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, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

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

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire set of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, Q_3, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$IQR = Q_3 – Q_1$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than** $(1.5)(\text{IQR})$ **below the first quartile or more than** $(1.5)(\text{IQR})$ **above the third quartile**. 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 2.3.1

For the following 13 real estate prices, calculate the IQR and determine if any prices are potential outliers. Prices are in dollars.
$389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000$

Answer

Order the data from smallest to largest.

$114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000$

$M = 488,800$

$Q_1 = \frac{230,500+387,000}{2} = 308,750$

$Q_3 = \frac{639,000+659,000}{2} = 649,000$

$IQR = 649,000–308,750 = 340,250$

$(1.5)(IQR) = (1.5)(340,250) = 510,375$

$Q_1–(1.5)(IQR) = 308,750–510,375 = –201,625$

$Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$

No house price is less than $–201,625$. However, $5,500,000$ is more than $1,159,375$. Therefore, $5,500,000$ is a potential **outlier**.

Example 2.3.2

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

Table 2.3.1

For the two data sets, find the following:

a. The interquartile range. Compare the two interquartile ranges.

b. Any outliers in either set.

Answer

a. The IQR for the day group is $Q_3–Q_1 = 82.5–56 = 26.5$

The IQR for the night group is $Q_3–Q_1 = 89–78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class IQR. This suggests more variation will be found in the day class's class test scores.

b. Day class outliers are found using the IQR times 1.5 rule. So,

- $Q_1 - IQR(1.5) = 56–26.5(1.5) = 16.25$
- $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 – IQR(1.5) = 78 – 11(1.5) = 61.5$
- $Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Example 2.3.3

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

Table 2.3.2

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Exercise 2.3.1

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

Table 2.3.3

Example 2.3.4

Using Table 2.3.2:

a. Find the 80th percentile.

b. Find the 90th percentile.

c. Find the first quartile. What is another name for the first quartile?

Answer

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$

b. The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.

c. Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

A Formula for Finding the kth Percentile

If you were to do a little research, you would find several formulas 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 points, or observations

- 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 data set. This is easier to understand in an example.

Example 2.3.5

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

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

1. Find the 70th percentile.
2. Find the 83rd percentile.

Answer

1.

- $k = 70$
- i = the index
- $n = 29$

$i = \frac{k}{100}(n+1) = \left(\frac{70}{100}\right)(29+1) = 21$. 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.

2.

- $k = 83^{rd}$ percentile
- i = the index
- $n = 29$

$i = \frac{k}{100}(n+1) = (\frac{83}{100})(29+1) = 24.9$, which is NOT an integer. Round it 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.

Exercise 2.3.2

Listed are 29 ages for Academy Award winning best actors *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.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- 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 for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x+0.5y}{n}(100)$. Then round to the nearest integer.

Example 2.3.6

Listed are 29 ages for Academy Award winning best actors *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

1. Find the percentile for 58.
2. Find the percentile for 25.

Answer

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

   $x = 18$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{18+0.5(1)}{29}(100) = 63.80$. 58 is the 64th percentile.

2. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

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

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 *k*th 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 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 2.3.7

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

Answer

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 finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example 2.3.8

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.

Answer

Seventy percent of students answered 16 or fewer questions correctly. Thirty percent of students answered 16 or more questions correctly. A higher percentile could be considered good, as answering more questions correctly is desirable.

Exercise 2.3.3

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.

Example 2.3.9

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

Answer

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Example 2.3.10

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 shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- $Q_1 = 20$
- Mdn = 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.

However, 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 calculate the five values, 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

This page titled 2.3: Measures of the Location of the Data is shared under a CC BY license and was authored, remixed, and/or curated by .

2.4: Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. Technically this is the arithmetic mean. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts meaning an equal number of observations on each side. The weight of 25 people are below this weight and 25 people are heavier than this weight. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set 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 letter used to represent the **sample mean** is an *x* with a bar over it (pronounced “x bar”): $\overline{x}$.

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

To see that both ways of calculating the mean are the same, consider the sample:
1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\overline{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

$$\overline{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. 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 2.4.1

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 mean and the median.

Answer

The calculation for the mean is:

$\overline{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$

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):

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

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

Example 2.4.2

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?

Answer

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

$M = 30,000$

(There are 49 people who earn $30,000 and one 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.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example 2.4.3

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.

Answer

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

Example 2.4.4

Five real estate exam scores are 430, 430, 480, 480, 495. The data set 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 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 qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Calculating the Arithmetic Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is 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: $mean = \frac{\text{data sum}}{\text{number of data values}}$ 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 $\frac{\text{lower boundary+upper boundary}}{2}$. We can now modify the mean definition to be $\textbf{Mean of Frequency Table} = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

Example 2.4.5

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

Table 2.4.1

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

Answer

Find the midpoints for all intervals

Table 2.4.2

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

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$ $53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$
- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

Exercise 2.4.1

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:

Table 2.4.3

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

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

This page titled 2.4: Measures of the Center of the Data is shared under a CC BY license and was authored, remixed, and/or curated by .

2.5: Sigma Notation and Calculating the Arithmetic Mean

Formula for Population Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Formula for Sample Mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Here are the formulas for a population mean and the sample mean. The Greek letter μ is the symbol for the population mean and $\overline{x}$ is the symbol for the sample mean. Both formulas have a mathematical symbol that tells us how to make the calculations. It is called Sigma notation because the symbol is the Greek capital letter sigma: Σ. Like all mathematical symbols it tells us what to do: just as the plus sign tells us to add and the x tells us to multiply. These are called mathematical operators. The Σ symbol tells us to add a specific list of numbers.

Let's say we have a sample of animals from the local animal shelter and we are interested in their average age. If we list each value, or observation, in a column, you can give each one an index number. The first number will be number 1 and the second number 2 and so on.

Table 2.5.1

Animal	Age
1	9
2	1
3	8.5
4	10.5
5	10
6	8.5
7	12
8	8
9	1
10	9.5

Each observation represents a particular animal in the sample. Purr is animal number one and is a 9 year old cat, Toto is animal number 2 and is a 1 year old puppy and so on.

To calculate the mean we are told by the formula to add up all these numbers, ages in this case, and then divide the sum by 10, the total number of animals in the sample.

Animal number one, the cat Purr, is designated as X_1, animal number 2, Toto, is designated as X_2 and so on through Dundee who is animal number 10 and is designated as X_{10}.

The *i* in the formula tells us which of the observations to add together. In this case it is X_1 through X_{10} which is all of them. We know which ones to add by the indexing notation, the $i = 1$ and the n or capital N for the population. For this example the indexing notation would be $i = 1$ and because it is a sample we use a small n on the top of the Σ which would be 10.

The standard deviation - which you'll learn about later in this chapter - requires the same mathematical operator.

The sum of the ages is found to be 78 and dividing by 10 gives us the sample mean age as 7.8 years.

This page titled 2.5: Sigma Notation and Calculating the Arithmetic Mean is shared under a CC BY license and was authored, remixed, and/or curated by .

2.6: Skewness and the Mean, Median, and Mode

Consider the following data set.
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

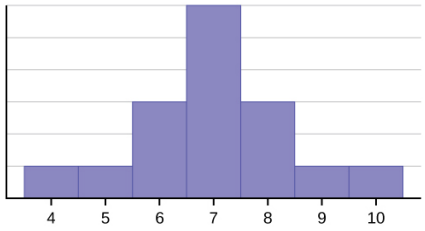


Figure 2.6.1

The histogram above 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 shape 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 seven 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 called **skewed to the left** because it is pulled out to the left.

The greater the deviation from zero indicates a greater degree of skewness. If the skewness is *negative* then the distribution is skewed left as in Figure 2.6.2.

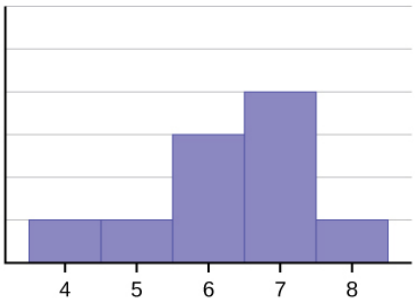


Figure 2.6.2

The mean is 6.3, the median is 6.5, and the mode is seven. **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 shown below for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is **skewed to the right**, or *positively* skewed.

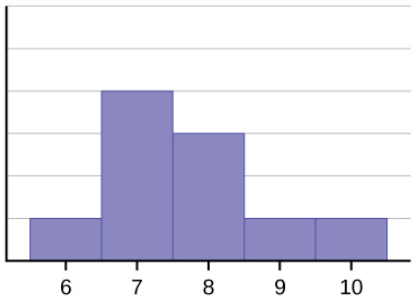


Figure 2.6.3

The mean is 7.7, the median is 7.5, and the mode is seven. 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, generally if the distribution of data is skewed to the left, the mean is 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.

The measure of skewness can be a negative number, and this is how we determine if the data are skewed right of left. The skewness for a normal distribution is zero, and any symmetric data should have skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. The skewness characterizes the degree of asymmetry of a distribution around its mean. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive X and a negative value signifies a distribution whose tail extends out towards more negative X. A zero measure of skewness will indicate a symmetrical distribution.

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

This page titled 2.6: Skewness and the Mean, Median, and Mode is shared under a CC BY license and was authored, remixed, and/or curated by .

2.7: Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean, on average.

The standard deviation

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

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

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

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

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

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

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

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

Formulas for the Sample Standard Deviation

- $s=\sqrt{\frac{\Sigma(x_i-\overline{x})^2}{n-1}}$ or $s=\sqrt{\frac{\Sigma f_i(x_i-\overline{x})^2}{n-1}}$
- For the sample standard deviation, the denominator is ***n* – 1**, that is the sample size minus 1.

Formulas for the Population Standard Deviation

- $\boldsymbol{\sigma}=\sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$ or $\sigma=\sqrt{\frac{\Sigma f_i(x_i-\mu)^2}{N}}$
- For the population standard deviation, the denominator is N, the number of items in the population.

In these formulas, f_i represents the frequency with which a given value of x (that is, x_i) appears. For example, if a value appears once, f_i is one for that value. If a different value appears three times in the data set or population, f_i for that value is three.

There are two important observations to note concerning the variance and standard deviation: the deviations are measured from the mean, and the deviations are squared. In principle, the deviations could be measured from any point, however, our interest is measurement from the center weight of the data, what is the most usual value of the observation. Later we will be trying to measure the "unusualness" of an observation or a sample mean and thus we need a measure from the mean. The second observation is that the deviations are squared. This does two things, first it makes the deviations all positive and second it changes the units of measurement from that of the mean and the original observations. If the data are weights then the mean is measured in pounds, but the variance is measured in pounds-squared. One reason to use the standard deviation is to return to the original units of measurement by taking the square root of the variance. Further, when the deviations are squared it explodes their value. For example, a deviation of 10 from the mean when squared is 100, but a deviation of 100 from the mean is 10,000. What this does is place great weight on outliers when calculating the variance.

Types of Variability in Samples

When trying to study a population, a sample is often used, either for convenience or because it is not possible to access the entire population. Variability is the term used to describe the differences that may occur in these outcomes. Common types of variability include the following:

- Observational or measurement variability
- Natural variability
- Induced variability
- Sample variability

Here are some examples to describe each type of variability.

Example 1: Measurement variability

Measurement variability occurs when there are differences in the instruments used to measure or in the people using those instruments. If we are gathering data on how long it takes for a ball to drop from a height by having students measure the time of the drop with a stopwatch, we may experience measurement variability if the two stopwatches used were made by different manufacturers. For example, one stopwatch measures to the nearest second, whereas the other one measures to the nearest tenth of a second. We also may experience measurement variability because two different people are gathering the data. Their reaction times in pressing the button on the stopwatch may differ; thus, the outcomes will vary accordingly. The differences in outcomes may be affected by measurement variability.

Example 2: Natural variability

Natural variability arises from the differences that naturally occur because members of a population differ from each other. For example, if we have two identical corn plants and we expose both plants to the same amount of water and sunlight, they may still grow at different rates simply because they are two different corn plants. The difference in outcomes may be explained by natural variability.

Example 3: Induced variability

Induced variability is the counterpart to natural variability; this occurs because we have artificially induced an element of variation (that, by definition, was not present naturally): For example, we assign people to two different groups to study memory, and we induce a variable in one group by limiting the amount of sleep they get. The difference in outcomes may be affected by induced variability.

Example 4: Sample variability

Sample variability occurs when multiple random samples are taken from the same population. For example, if I conduct four surveys of 50 people randomly selected from a given population, the differences in outcomes may be affected by sample variability.

Example 2.7.1

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

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;

$$\overline{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

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

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

Table 2.7.1

Data	Freq.	Deviations	Deviations^2	(Freq.)(Deviations^2)
x	f	$(x - \overline{x})$	$(x–\overline{x})^2$	$(f)(x–\overline{x})^2$
9	1	$9–10.525 = –1.525$	$(–1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5–10.525 = –1.025$	$(–1.025)2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10–10.525 = –0.525$	$(–0.525)2 = 0.275625$	$4 \times 0.275625 = 1.1025$
10.5	4	$10.5–10.525 = –0.025$	$(–0.025)2 = 0.000625$	$4 \times 0.000625 = 0.0025$
11	6	$11–10.525 = 0.475$	$(0.475)2 = 0.225625$	$6 \times 0.225625 = 1.35375$
11.5	3	$11.5–10.525 = 0.975$	$(0.975)2 = 0.950625$	$3 \times 0.950625 = 2.851875$
				The total is 9.7375

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 $(20–1)$:

$s^2 = \frac{9.7375}{20–1} = 0.5125$

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, $s = 0.72$.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For Example 2.7.1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation. By squaring the deviations we are placing an extreme penalty on observations that are far from the mean; these observations get greater weight in the calculations of the variance. We will see later on that the variance (standard deviation) plays the critical role in determining our conclusions in inferential statistics. We can begin now by using the standard deviation as a measure of "unusualness." "How did you do on the test?" "Terrific! Two standard deviations above the mean." This, we will see, is an unusually good exam grade.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n–1 = 20–1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one $(n–1)$. Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** This estimate requires us to use an estimate of the population mean rather than the actual population mean. Based on the theoretical mathematics that lies behind these calculations, dividing by $(n–1)$ gives a better estimate of the population variance.

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

Example 2.7.2

Use the following data (first exam scores) from Dr. Doom'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

Answer

a. See Table 2.7.2 below

b.

i. The sample mean = 73.5
ii. The sample standard deviation = 17.9
iii. The median = 73
iv. The first quartile = 61
v. The third quartile = 90
vi. $IQR = 90–61 = 29$

Table 2.7.2

Data	Frequency	Relative frequency	Cumulative relative frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1? Answer: Rounding)

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value x, calculate how many standard deviations away from its mean the value is.
- Use the formula: x = mean + (#of STDEVs)(standard deviation); solve for #of STDEVs.
- $\#\text{ of } STDEVs = \frac{x - \text{ mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#of STDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Table 2.7.3

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

Example 2.7.3

Two students, John and Ali, from different high schools, wanted to find out who had the higher GPA when compared to his school. Which student had the higher GPA compared to his own school?

Table 2.7.4

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

Answer

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

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

For John, $z = \#\text{ ofSTDEV } s = \frac{2.85 \cdot 3.0}{0.7} = -0.21$

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

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** his 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.

Exercise 2.7.1

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

Table 2.7.5

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

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.

- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a normal distribution, which we will examine in great detail later:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the **Empirical Rule**.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

This page titled 2.7: Measures of the Spread of the Data is shared under a CC BY license and was authored, remixed, and/or curated by .

2.8: Chapter 2 Key Terms

Frequency

the number of times a value of the data occurs

Frequency Table

a data representation in which grouped data is displayed along with the corresponding frequencies

Histogram

a graphical representation in *x*-*y* form of the distribution of data in a data set; *x* represents the data and *y* represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Interquartile Range

or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Mean (arithmetic)

a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x}= \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu= \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$

Median

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 than the median. The median may or may not be part of the data.

Midpoint

the mean of an interval in a frequency table

Mode

the value that appears most frequently in a set of data

Outlier

an observation that does not fit the rest of the data

Percentile

a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation.

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \overline{x}$ where *x* is a value of the data and $\overline{x}$ is the sample mean. The sample variance is equal to the sum of the

squares of the deviations divided by the difference of the sample size and one.

This page titled 2.8: Chapter 2 Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by .

2.9: Chapter 2 Review

2.1 Display Data

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

2.2 Measures of the Location of the Data

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other observations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile,the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the the 75th percentile. The interquartile range, or IQR, is the range of the middle 50 percent of the data values. The IQR is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 – IQR(1.5)$

2.3 Measures of the Center of the Data

The **mean** and the **median** can be calculated to help you find the "center" of a data set. The mean is the most commonly used estimate of central tendency, but the median is the best measurement when a data set contains several outliers or extreme values. The **mode** will tell you the most frequently occurring datum (or data) in your data set.

The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

2.6 Skewness and the Mean, Median, and Mode

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. In a **right (or positively) skewed** distribution, the mean will be higher than the median. In a **left (or negatively) skewed** distribution, the mean will be lower than the median. In a **normally shaped distribution**, the mean and median will be (approximately) equal.

2.7 Measures of the Spread of the Data

The standard deviation can be used to represent the spread of data, and it is a measure of how far, on average, all of the individual scores are from the mean.

There are different equations to use if we are calculating the standard deviation of a sample or of a population.

- $s = \sqrt{\frac{\Sigma(x_i - \overline{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f_i(x_i - \overline{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula is $\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f_i(x_i - \mu)^2}{N}}$

This page titled 2.9: Chapter 2 Review is shared under a CC BY license and was authored, remixed, and/or curated by .

2.10: Chapter 2 Formula Review

2.2 Measures of the Location of the Data

$i = \left(\frac{k}{100}\right)(n+1)$

where i = the ranking or position of a data value,

k = the kth percentile,

n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x+0.5y}{n}\right)(100)$

where x = the number of 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 for which you want to find the percentile,

n = total number of data

2.3 Measures of the Center of the Data

$\mu = \frac{\sum fm}{\sum f}$ Where f = interval frequencies and m = interval midpoints.

The arithmetic mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$

The arithmetic mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$

2.7 Measures of the Spread of the Data

$s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2}$ where $s_x =$ sample standard deviation, $\overline{x} =$ sample mean

Formulas for Sample Standard Deviation $s = \sqrt{\frac{\Sigma(x_i-\overline{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f_i(x_i-\overline{x})^2}{n-1}}$

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

Formulas for Population Standard Deviation $\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f_i(x_i-\mu)^2}{N}}$

For the population standard deviation, the denominator is *N*, the number of items in the population.

This page titled 2.10: Chapter 2 Formula Review is shared under a CC BY license and was authored, remixed, and/or curated by .

2.11: Chapter 2 Homework

2.2 Display Data

For the next three exercises, use the data to construct an appropriate graph.

1. 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

Table 2.11.1

2. 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

Table 2.11.2

3. 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
1	18
2	36
3	7
4	2

Table 2.11.3

4. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. Table 2.11.4 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%
Winter	6	18%

Table 2.11.4

5. Using the data from Mrs. Ramirez's math class supplied in the previous question, construct a bar graph showing the percentages.

6. David County has six high schools. Each school sent students to participate in a county-wide science competition. Table 2.11.5 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%

Table 2.11.5

7. Use the data from the David County science competition supplied in Table 2.11.5. Construct a bar graph that shows the county-wide population percentage of students at each school.

8. 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; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete Table 2.11.6.

Data value (# cars)	Frequency	Relative frequency	Cumulative relative frequency

Table 2.11.6

9. What does the frequency column in Table 2.11.6 sum to? Why?

10. What does the relative frequency column in Table 2.11.6 sum to? Why?

11. What is the difference between relative frequency and frequency for each data value in Table 2.11.6?

12. What is the difference between cumulative relative frequency and relative frequency for each data value?

13. To construct the histogram for the data in Table 2.11.6, 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.



Figure 2.11.1

14. Construct a frequency polygon for the following:

Table 2.11.7

a.

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

Table 2.11.8

b.

Actual speed in a 30 MPH zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

Table 2.11.9

c.

Tar (mg) in nonfiltered cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

15. Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Table 2.11.10

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

16. Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed 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?

Table 2.11.11

Life expectancy at birth – women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Table 2.11.12

Life expectancy at birth – men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

17. 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.

Table 2.11.13

Sex/Year	**1855**	**1856**	**1857**	**1858**	**1859**	**1860**	**1861**
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Table 2.11.14

Sex/Year	**1862**	**1863**	**1864**	**1865**	**1866**	**1867**	**1868**	**1869**
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Table 2.11.15

Sex/Year	**1870**	**1871**	**1872**	**1873**	**1874**	**1875**
Female	56,431	56,099	57,472	58,233	60,109	60,146
Male	58,959	60,029	61,293	61,467	63,602	63,432
Total	115,390	116,128	118,765	119,700	123,711	123,578

18. The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Table 2.11.16

Year	**1961**	**1962**	**1963**	**1964**	**1965**	**1966**	**1967**
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Table 2.11.17

Year	**1968**	**1969**	**1970**	**1971**	**1972**	**1973**
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

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.

2.3 Measures of the Location of the Data

19. Listed are 29 ages for Academy Award winning best actors *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.

20. Listed are 32 ages for Academy Award winning best actors *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.

21. Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

22.

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. 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.

23.

a. For runners in a race, a higher speed means a faster run. 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.

24. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

25. 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.

26. 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.

27. 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 had $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 sentence that interprets the 90th percentile in the context of this problem.

28. 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"?

29. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Use the following information to answer the next six exercises. 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; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

30. First quartile = _____

31. Second quartile = median = 50th percentile = _____

32. Third quartile = _____

33. Interquartile range (*IQR*) = _____ – _____ = _____

34. 10th percentile = _____

35. 70th percentile = _____

2.4 Measures of the Center of the Data

36. Find the approximate 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

Table 2.11.18

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

Table 2.11.19

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

Table 2.11.20

Use the following information to answer the next three exercises: 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

37. Calculate the mean.

38. Identify the median.

39. Identify the mode.

Use the following information to answer the next three exercises: 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; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Calculate the following:

40. sample mean = _____

41. median = _____

42. mode = _____

2.5 Sigma Notation and Calculating the Arithmetic Mean

43. A group of 10 children are on a scavenger hunt to find different color rocks. The results are shown in the Table 2.11.21 below. The column on the right shows the number of colors of rocks each child has. What is the mean number of rocks?

Child	Rock colors
1	5
2	5
3	6
4	2
5	4
6	3
7	7
8	2
9	1
10	10

Table 2.11.21

44. A group of children are measured to determine the average height of the group. The results are in Table 2.11.22 below. What is the mean height of the group to the nearest hundredth of an inch?

Child	Height in inches
Adam	45.21
Betty	39.45
Charlie	43.78
Donna	48.76
Earl	37.39
Fran	39.90
George	45.56
Heather	46.24

Table 2.11.22

45. A person compares prices for five automobiles. The results are in Table 2.11.23. What is the mean price of the cars the person has considered?

Price
$20,987
$22,008
$19,998
$23,433
$21,444

Table 2.11.23

46. A customer protection service has obtained 8 bags of candy that are supposed to contain 16 ounces of candy each. The candy is weighed to determine if the average weight is at least the claimed 16 ounces. The results are in given in Table 2.11.24. What is the mean weight of a bag of candy in the sample?

Weight in ounces
15.65
16.09
16.01
15.99
16.02
16.00
15.98
16.08

Table 2.11.24

47. A teacher records grades for a class of 70, 72, 79, 81, 82, 82, 83, 90, and 95. What is the mean of these grades?

48. A family is polled to see the mean of the number of hours per day the television set is on. The results, starting with Sunday, are 6, 3, 2, 3, 1, 3, and 7 hours. What is the average number of hours the family had the television set on to the nearest whole number?

49. A city received the following rainfall for a recent year. What is the mean number of inches of rainfall the city received monthly, to the nearest hundredth of an inch? Use Table 2.11.25.

Month	Rainfall in inches
January	2.21
February	3.12
March	4.11
April	2.09
May	0.99
June	1.08
July	2.99
August	0.08
September	0.52
October	1.89
November	2.00
December	3.06

Table 2.11.25

50. A football team scored the following points in its first 8 games of the new season. Starting at game 1 and in order the scores are 14, 14, 24, 21, 7, 0, 38, and 28. What is the mean number of points the team scored in these eight games?

51. What is the mean and standard deviation of the data set given? 5, 10, 20

52. What is the mean and standard deviation of the data set given? 9.000, 15.00, 21.00

53. What is the mean of the data set given? 7.0, 10.0, 39.2

54. What is the mean and standard deviation of the data set given? 17.00, 10.00, 19.00

55. What is the average and standard deviation for the values that follow? 1.0, 2.0, 1.5

56. What is the average and standard deviation for the values that follow? 0.80, 2.0, 5.0

57. What is the average and standard deviation for the values that follow? 0.90, 1.1, 1.2

58. What is the average and standard deviation for the values that follow? 4.2, 4.3, 4.5

2.6 Skewness and the Mean, Median, and Mode

Use the following instructions to answer the next three exercises. State whether the data are symmetrical, skewed to the left, or skewed to the right.

59. 1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5

60. 16; 17; 19; 22; 22; 22; 22; 22; 23

61. 87; 87; 87; 87; 87; 88; 89; 89; 90; 91

62. When the data are skewed left, what is the typical relationship between the mean and median?

63. When the data are symmetrical, what is the typical relationship between the mean and median?

64. What word describes a distribution that has two modes?

65. Describe the shape of the distribution in Figure 2.11.2.

66. Describe the relationship between the mode and the median of the distribution in Figure 2.11.2

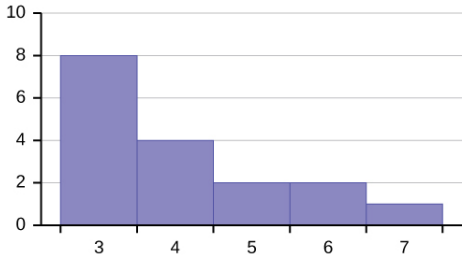


Figure 2.11.2

67. Describe the relationship between the mean and the median of the distribution in Figure 2.11.3.

68. Describe the shape of the distribution in Figure 2.11.3.

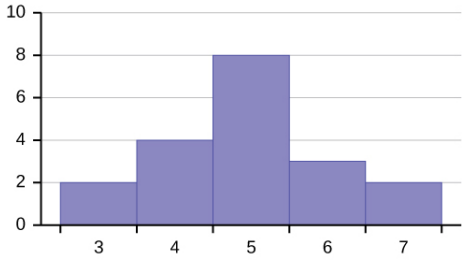


Figure 2.11.3

69. Describe the relationship between the mode and the median of the distribution in Figure 2.11.3.

70. Are the mean and the median the exact same in the distribution in Figure 2.11.3? Why or why not?

71. Describe the shape of the distribution in Figure 2.11.4.

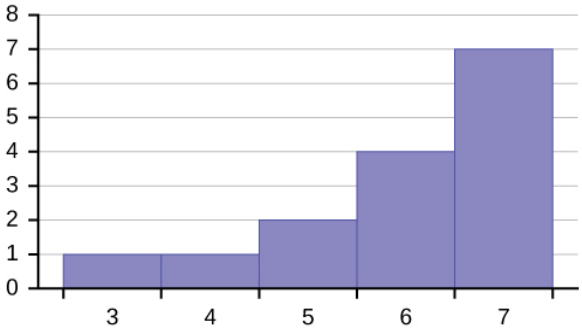


Figure 2.11.4

72. Describe the relationship between the mode and the median of the distribution in Figure 2.11.4.

73. Describe the relationship between the mean and the median of the distribution in Figure 2.11.4.

74. The mean and median for the following data are the same. Is the data perfectly symmetrical? Why or why not?

3; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7

75. Which is the greatest, the mean, the mode, or the median of the following data set?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

76. Which is the least, the mean, the mode, and the median of the following data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

77. Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

78. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

2.7 Measures of the Spread of the Data

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

79. Find the standard deviation and round to the nearest tenth.

80. Find the value that is one standard deviation below the mean.

81. Two baseball players, Fredo and Karl, 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?

Table 2.11.26

Baseball player	Batting average	Team batting average	Team standard deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

82. Use Table 2.11.26 to find the value that is three standard deviations:

- above the mean
- below the mean

83. Find the approximate mean for the following frequency tables using the formula.

Table 2.11.27

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

Table 2.11.28

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

Table 2.11.29

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

2.2 Display Data

84. Table 2.11.30 contains the 2019 poverty rates in U.S. states and Washington, DC.

Table 2.11.30

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	15.6	Kentucky	16.0	North Dakota	10.5
Alaska	10.2	Louisiana	18.8	Ohio	13.0
Arizona	13.5	Maine	10.9	Oklahoma	15.1
Arkansas	16.0	Maryland	9.1	Oregon	11.5
California	11.8	Massachusetts	9.5	Pennsylvania	12
Colorado	9.4	Michigan	12.9	Rhode Island	11.6
Connecticut	9.9	Minnesota	8.9	South Carolina	13.9
Delaware	11.2	Mississippi	19.5	South Dakota	11.9
Washington, DC	14.1	Missouri	12.9	Tennessee	13.8
Florida	12.7	Montana	12.6	Texas	13.6
Georgia	13.5	Nebraska	9.9	Utah	8.8
Hawaii	9.0	Nevada	12.7	Vermont	10.1
Idaho	11.0	New Hampshire	7.5	Virginia	9.9
Illinois	11.4	New Jersey	9.1	Washington	9.8
Indiana	11.9	New Mexico	17.5	West Virginia	16.2
Iowa	11.0	New York	13.1	Wisconsin	10.4
Kansas	11.3	North Carolina	13.6	Wyoming	9.9

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the poverty 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."

85. 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:

Table 2.11.31 Publisher A

# of books	Freq.	Rel. freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Table 2.11.32 Publisher B

# of books	Freq.	Rel. freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.11.33 Publisher C

# of books	Freq.	Rel. freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

a. Find the relative frequencies for each survey. Write them in the charts.
b. 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.

86. 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. Following is a summary of the bills for each group.

Table 2.11.34 Singles

Amount($)	Frequency	Rel. frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Table 2.11.35 Couples

Amount($)	Frequency	Rel. frequency
100–150	5	
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	

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 for the couples by hand. 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 individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

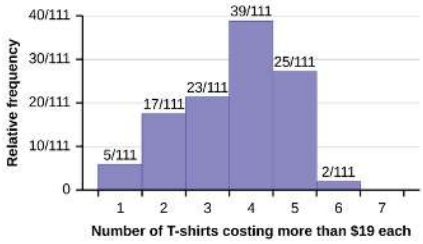
87. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

Table 2.11.36

# of movies	Frequency	Relative frequency	Cumulative relative frequency
0	5		
1	9		
2	6		
3	4		
4	1		

a. Construct a histogram of the data.
b. Complete the columns of the chart.
c. Find the median, and interpret it.
d. Find the mode, and interpret it.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.



88. 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

89. 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

90. Following are the 2019 poverty rates by U.S. states and Washington, DC.

Table 2.11.37

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	15.6	Kentucky	16.0	North Dakota	10.5
Alaska	10.2	Louisiana	18.8	Ohio	13.0
Arizona	13.5	Maine	10.9	Oklahoma	15.1
Arkansas	16.0	Maryland	9.1	Oregon	11.5
California	11.8	Massachusetts	9.5	Pennsylvania	12
Colorado	9.4	Michigan	12.9	Rhode Island	11.6
Connecticut	9.9	Minnesota	8.9	South Carolina	13.9
Delaware	11.2	Mississippi	19.5	South Dakota	11.9
Washington, DC	14.1	Missouri	12.9	Tennessee	13.8
Florida	12.7	Montana	12.6	Texas	13.6
Georgia	13.5	Nebraska	9.9	Utah	8.8
Hawaii	9.0	Nevada	12.7	Vermont	10.1
Idaho	11.0	New Hampshire	7.5	Virginia	9.9
Illinois	11.4	New Jersey	9.1	Washington	9.8
Indiana	11.9	New Mexico	17.5	West Virginia	16.2
Iowa	11.0	New York	13.1	Wisconsin	10.4
Kansas	11.3	North Carolina	13.6	Wyoming	9.9

Construct a bar graph of poverty rates of your state and the four states closest to your state. Hint: Label the *x*-axis with the states.

2.3 Measures of the Location of the Data

91. The median age for African Americans in the U.S. currently is 34 years; for Caucasians in the U.S., it is 43 years. How might it be possible for African Americans and Caucasians to die at approximately the same age, but for the median ages to differ?

92. Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in Table 2.11.38. Also, include left endpoint, but not the right endpoint.

Table 2.11.38

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

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.
   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

2.4 Measures of the Center of the Data

93. The hottest temperatures on record by country range from 86.2°F (Greenland) to 134.0°F (the United States). This data is summarized in the following table.

Table 2.11.39

Hottest temperature on record (°F)	Number of countries
79.5–89.5	2
89.5–99.5	13
99.5–109.5	39
109.5–119.5	41
119.5–129.5	28
129.5–139.5	2

a. What is the best estimate of the average hottest temperature for all countries?
b. The United States has a hottest temperature recording of 134.0°F. Is this temperature above average or below? How does the United States compare to other countries?

94. Table 2.11.40 gives the percent of children under five earning various score ranges on a motor skills assessment. What is the best estimate for the mean score of all these children on the assessment?

Table 2.11.40

Score range on assessment	Number of children
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

95. A sample of 10 prices is chosen from a population of 100 similar items. The values obtained from the sample, and the values for the population, are given in Table 2.11.41 and Table 2.11.42, respectively.

a. Is the mean of the sample within $1 of the population mean?
b. What is the difference in the sample and population means?
c. Interpret the sample mean.
d. Interpret the population mean.
e. Calculate and interpret the standard deviation of the sample.

f. Calculate and interpret the standard deviation of the population.

Table 2.11.41

Prices of the sample
$21
$23
$21
$24
$22
$22
$25
$21
$20
$24

Table 2.11.42

Prices of the population	Frequency
$20	20
$21	35
$22	15
$23	10
$24	18
$25	2

96. A standardized test is given to ten people at the beginning of the school year with the results given in Table 2.11.43 below. At the end of the year the same people were again tested.

a. What is the average improvement?
b. Does it matter if the means are subtracted, or if the individual values are subtracted?

Table 2.11.43

Student	Beginning score	Ending score
1	1100	1120
2	980	1030
3	1200	1208
4	998	1000
5	893	948
6	1015	1030
7	1217	1224
8	1232	1245
9	967	988
10	988	997

97. A small class of 7 students has a mean grade of 82 on a test.

a. If six of the grades are 80, 82, 86, 90, 90, and 95, what is the other grade?
b. Interpret the mean in words.
c. Find the median, and interpret it.
d. Find the mode, and interpret it.

98. A class of 20 students has a mean grade of 80 on a test. Nineteen of the students has a mean grade between 79 and 82, inclusive.

a. What is the lowest possible grade of the other student?
b. What is the highest possible grade of the other student?

99. If the mean of 20 prices is $10.39, and 5 of the items with a mean of $10.99 are sampled, what is the mean of the other 15 prices?

2.6 Skewness and the Mean, Median, and Mode

105. The median age of the U.S. 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. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

2.7 Measures of the Spread of the Data

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- $\mu = 1000$ FTES
- median $= 1,014$ FTES
- $\sigma = 474$ FTES
- first quartile $= 528.5$ FTES
- third quartile $= 1,447.5$ FTES
- $n = 29$ years

106. A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

107. 75% of all years have an FTES:

a. at or below: _____
b. at or above: _____

108. The population standard deviation = _____

109. What percent of the FTES were from 528.5 to 1447.5? How do you know?

110. What is the IQR? What does the IQR represent?

111. 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.

Table 2.11.44

Year	**2005–06**	**2006–07**	**2007–08**	**2008–09**	**2009–10**	**2010–11**
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

112. Calculate the mean, median, standard deviation, the first quartile, the third quartile and the IQR. Round to one decimal place.

113. 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?

114. Three students were applying to the same graduate school. They came from schools with different grading systems, but we will assume that each school's GPAs follow a normal distribution.

a. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

b. Interpret the mean and standard deviation of GPAs at Thuy's school.

c. If we assume that Thuy's school has a maximum GPA of 4.0, what shape is the distribution of GPAs at Thuy's school likely to be?

Table 2.11.45

Student	GPA	School Mean GPA	School SD GPA
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

115. 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.

116. 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 1 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 his or her class? Explain why.

117. The hottest temperatures on record by country range from 86.2°F (Greenland) to 134.0°F (the United States). This data is summarized in Table 2.11.46.

Hottest temperature on record (°F)	Number of countries
79.5–89.5	2
89.5–99.5	13
99.5–109.5	39
109.5–119.5	41
119.5–129.5	28
129.5–139.5	2

Table 2.11.46

What is the best estimate of the average hottest recorded temperature for these countries? How "unusual" is the United States' hottest temperature of 134.0 °F compared to the average here? Explain.

118. Table 2.11.47 gives the percent of children under five earning various score ranges on a motor skills assessment. What is the best estimate for the most common score of all these children on the assessment? How did you determine this?

Table 2.11.47

Score range on assessment	Number of children
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

119. 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.

Table 2.11.47

	Javier	Ercilia
$\overline{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

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 two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

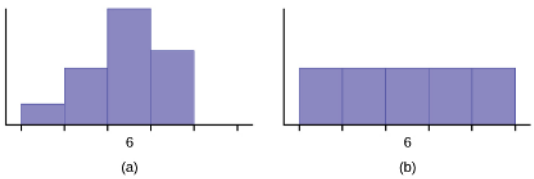


Figure 2.11.5

Use the following information to answer the next three exercises: 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.

Table 2.11.48

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

120. What is the IQR?

a. 8
b. 11
c. 15
d. 35

121. What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

122. Is this a sample or the entire population?

a. sample
b. entire population
c. neither

123. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

Table 2.11.49

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

a. Find the sample mean $\overline{x}$.
b. Find the approximate sample standard deviation, s.

124. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

Table 2.11.50

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean $\overline{x}$, rounded to one decimal place.
b. Find the sample standard deviation, s.
c. Construct a histogram of the data.
d. Add relative frequency and cumulative relative frequency columns to the chart.
e. Find the first quartile.
f. Find the median.
g. Find the IQR.
h. Find the mode.
i. Interpret the sample mean, standard deviation, Q1, median, Q3, IQR, and mode.
j. What percent of the students owned at least five pairs?
k. Find the 40th percentile.
l. Find the 90th percentile.
m. Construct a line graph of the data.
n. Construct a stemplot of the data.

125. 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. The middle 50% of the weights are from _____ to _____.
f. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
g. 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?
h. Assume the population was the San Francisco 49ers. Find:
   i. the population mean, μ.
   ii. the population standard deviation, $sigma$.
   iii. the weight that is two standard deviations below the mean.
   iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
i. 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?

126. 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 sample?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

127. Refer to Figure 2.11.6 determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

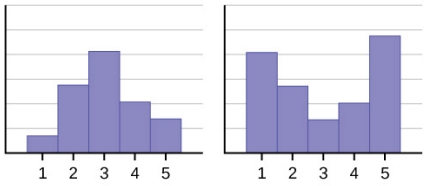


Figure 2.11.6

a. The medians for both graphs are the same.
b. We cannot determine if any of the means for both 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 both graphs is different.

128. 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 = 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. The middle 50% of the conferences last from _____ days to _____ days.
f. Calculate the sample mean of days of engineering conferences.
g. Calculate the sample standard deviation of days of engineering conferences.
h. Find the mode.
i. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
j. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

129. A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

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 8000 would be how many standard deviations away from the mean?

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

Table 2.11.51

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

130. The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

131. The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

132. 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 Table 2.11.52.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.11.52

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?

This page titled 2.11: Chapter 2 Homework is shared under a CC BY license and was authored, remixed, and/or curated by .

2.12: Chapter 2 Solutions

1.



Figure 2.12.1

3.

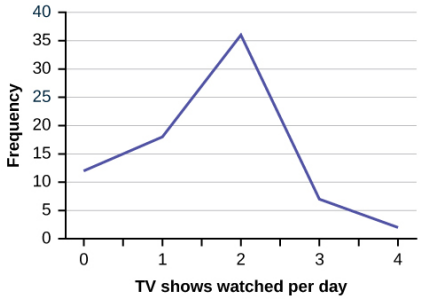


Figure 2.12.2

5.

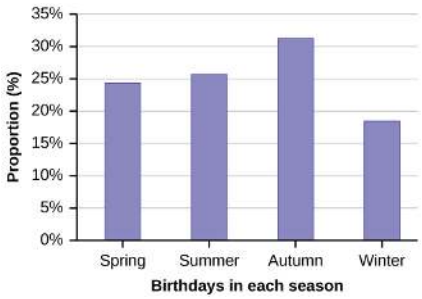


Figure 2.12.3

7.

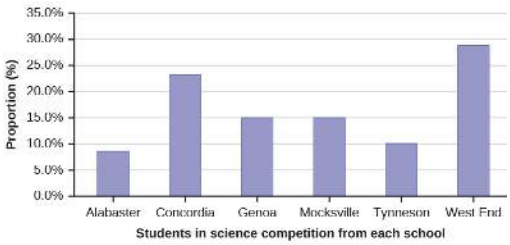


Figure 2.12.4

9. 65

11. The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

13. Answers will vary. One possible histogram is shown:

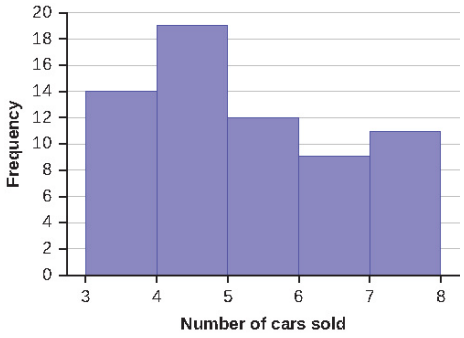


Figure 2.12.5

15. 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.

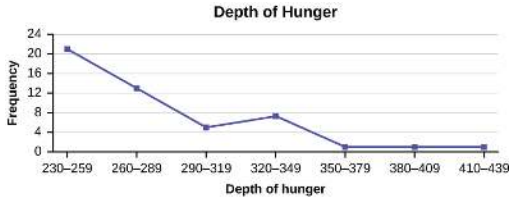


Figure 2.12.6

17.

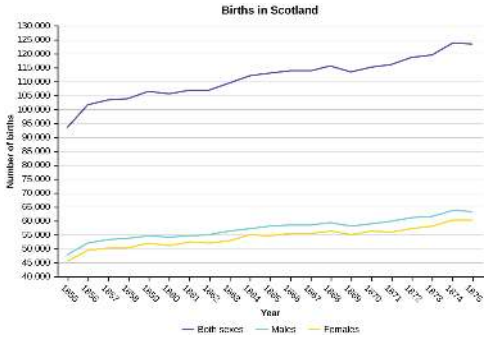


Figure 2.12.7

19.

a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

21. Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

$x = 143$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{143+0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

23.

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.
b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

25. When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

27. The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

29. You can afford 34% of houses. 66% of the houses are too expensive for your budget. Interpretation: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

31. 4

33. $6–4 = 2$

35. 6

37. Mean: $\frac{738}{27} = 27.33$

39. The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

41. 4

44. 39.48 in.

45. $21,574

46. 15.98 ounces

47. 81.56

48. 4 hours

49. 2.01 inches

50. $\frac{146}{8} = 18.25$

51. $\overline{x} = \frac{35}{3} = 11.67$; $s = 7.64$

52. $\overline{x} = \frac{45}{3} = 15.00$; $s = 6.00$

53. $\overline{x} = \frac{56.2}{3} = 18.73$; $s = 17.79$

54. $\overline{x} = \frac{46}{3} = 15.33$; $s = 4.73$

55. $\overline{x} = \frac{4.5}{3} = 1.50$; $s = 0.50$

56. $\overline{x} = \frac{7.8}{3} = 2.60$; $s = 2.16$

57. $\overline{x} = \frac{3.2}{3} = 1.07$; $s = 0.15$

58. $\overline{x} = \frac{13}{3} = 4.33$; $s = 0.15$

59. The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

61. The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

63. When the data are symmetrical, the mean and median are close or the same.

65. The distribution is skewed right because it looks pulled out to the right.

67. The mean is 4.9 and is about the same as the median, which is 5.

69. The mode and the median are the same. In this case, they are both 5.

71. The distribution is skewed left because it looks pulled out to the left.

73. The mean and the median are both 6.

75. The mode is 12, the median is 12.5, and the mean is 15.1. The mean is the largest.

77. The mean tends to reflect skewing the most because it is affected the most by outliers.

79. $s = 34.5$

81.

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.

83.

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$
b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{38045.3}{101} - 60.94^2} = 7.62$
c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

84. Answers will vary.

86.

Amount($)	Frequency	Relative frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Table 2.12.1 Singles

Amount($)	Frequency	Relative frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

Table 2.12.2 Couples

a. See Table 2.12.1 and Table 2.12.2.
b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).

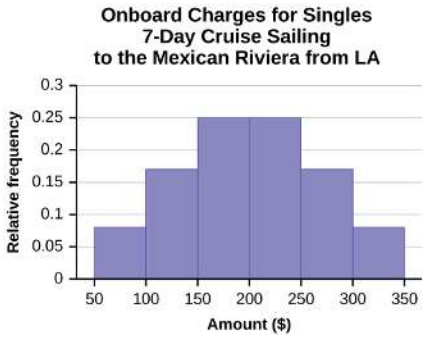


Figure 2.12.8

c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).



Figure 2.12.9

d. Compare the two graphs:

i. Answers may vary. Possible answers include:

- Both graphs have a single peak.
- Both graphs use class intervals with width equal to $50.

ii. Answers may vary. Possible answers include:

- The couples graph has a class interval with no values.
- It takes almost twice as many class intervals to display the data for couples.

iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.

f. Compare the graph for the Singles with the new graph for the Couples:

i.
- Both graphs have a single peak.
- Both graphs display 6 class intervals.
- Both graphs show the same general pattern.

ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.

h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

88. c

90. Answers will vary.

92.

a. $1-(0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01)=0.06$

b. $0.19+0.26+0.18=0.63$

c. Check student's solution.
d. 40th percentile will fall between 30,000 and 40,000

80th percentile will fall between 50,000 and 75,000

e. Check student's solution.

94. The mean score, $\overline{x} = \frac{1328.65}{50} = 26.57$

95.

a. Yes
b. The sample is $0.23 higher.
c. The average sample price was $22.
d. The average population price was $21.77.
e. The average difference of all the prices in the sample from the average was $1.63.
f. On average, all of the prices in the population differed from their average by $1.21.

96.

a. 20
b. No

97.

a. 51
b. On average, students in this class scored 82 on the test.
c. In order, the seven students scored: 51, 80, 82, 86, 90, 90, and 95. The middlemost score here is 86, so this is the median. Possible interpretation: 50% of students scored below an 86 on this test.
d. 90. The most commonly occurring test score was 90.

98.

a. 42
b. 99

99. $10.19

106. The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

108. 474 FTES

110. 919

112.

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- $IQR = 245$

113. Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

115. For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

117.

- $\overline{x} = \frac{13922.50}{125} = 111.38$
- The hottest temperature on record in the United States is considerably higher (22.62 °F) than the average hottest temperature across all countries.

120. a

122. b

123.

a. 1.48
b. 1.12

125.

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
b. 241
c. 205.5
d. 272.5
e. 205.5, 272.5
f. sample
g. population
h. i. 236.34
ii. 37.50
iii. 161.34
iv. 0.84 std. dev. below the mean
i. Young

127.

a. True
b. True
c. True
d. False

129.

a.

Table 2.12.3

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student's solution.
c. mode
d. 8628.74
e. 6943.88

f. –0.09

131. a

This page titled 2.12: Chapter 2 Solutions is shared under a CC BY license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

3: Probability

3.1: Introduction to Probability
3.2: Probability Terminology
3.3: Independent and Mutually Exclusive Events
3.4: Two Basic Rules of Probability
3.5: Contingency Tables and Probability Trees
3.6: Chapter 3 Key Terms
3.7: Chapter 3 Review
3.8: Chapter 3 Formula Review
3.9: Chapter 3 Homework
3.10: Chapter 3 Solutions
3.11: Chapter 3 References

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

3.1: Introduction to Probability

It is often necessary to "guess" 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 play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.



Figure 3.1.1 Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

You have, more than likely, 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.

3.1: Introduction to Probability is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

- **3.0: Introduction to Probability** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.
- **Current page** is licensed CC BY 4.0.

3.2: Probability Terminology

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. 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: to list the possible outcomes, to create a tree diagram, or to create 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 and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means the 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).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (*H*) and a Tail (*T*) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where $T =$ tails and $H =$ heads. The sample space has four outcomes. Let $A =$ getting one head. There are two outcomes that meet this condition $\{HT, TH\}$, so $P(A) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers $\{1, 2, 3, 4, 5, 6\}$ on its faces. Let event $E =$ be rolling a number that is at least five. There are two outcomes $\{5, 6\}$, so $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, 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.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

"∪" **Event: The Union**

An outcome is in the event $A \cup B$ if the outcome is in A or is in B or is in both A and B. For example, let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. $A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

"∩" **Event: The Intersection (Joint Probability)**

An outcome is in the event $A \cap B$ if the outcome is in both A and B at the same time. For example, let A and B be $\{1, 2, 3, 4, 5\}$ and $\{4, 5, 6, 7, 8\}$, respectively. Then $A \cap B = \{4, 5\}$.

The **complement** of event A is denoted A' (read out loud as "A prime"). A' consists of all outcomes that are **NOT** in A. Notice that $P(A) + P(A') = 1$. For example, let $S = \{1, 2, 3, 4, 5, 6\}$ and let $A = \{1, 2, 3, 4\}$. Then, we will have $A' = \{5, 6\}$, giving us $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and $P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$.

The **conditional probability** of A given B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space.** We calculate the probability of A from the reduced sample space B. The formula to calculate $P(A|B)$ is $P(A|B) = \frac{P(A \cap B)}{P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$. Let $A =$ face is 2 or 3 and $B =$ face is even $(2, 4, 6)$. To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space $B = \{2, 4, 6\}$. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$$P(A|B) = \frac{\frac{(\text{ number of outcomes that are 2 or 3 and even in } S)}{6}}{\frac{(\text{ number of outcomes that are even in } S)}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Odds

The odds of an event presents the probability as a ratio of success to failure. This is common in various gambling formats. Mathematically, the odds of an event can be defined as:

$$\frac{P(A)}{1 - P(A)}$$

where $P(A)$ is the probability of success and of course $1 - P(A)$ is the probability of failure. Odds are always quoted as "numerator to denominator," e.g. 2 to 1. Here the probability of winning is twice that of losing; thus, the probability of winning is 0.66. A probability of winning of 0.60 would generate odds in favor of winning of 3 to 2.

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example 3.2.1

The sample space S is the whole numbers starting at one and less than 20.

1. $S =$ _____

Let event A = the even numbers and event B = numbers greater than 13.

2. $A =$ _____, $B =$ _____
3. $P(A) =$ _____, $P(B) =$ _____
4. $A \cap B =$ _____, $A \cup B =$ _____
5. $P(A \cap B) =$ _____, $P(A \cup B) =$ _____
6. $A' =$ _____, $P(A') =$ _____
7. $P(A) + P(A') =$ _____
8. $P(A|B) =$ _____, $P(B|A) =$ _____; are the probabilities equal?

Answer

1. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
2. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$

3. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$
4. $A \cap B = \{14, 16, 18\}$, $A \cup B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
5. $P(A \cap B) = \frac{3}{19}$, $P(A \cup B) = \frac{12}{19}$
6. $A' = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$, $P(A') = \frac{10}{19}$
7. $P(A) + P(A') = \frac{9}{19} + \frac{10}{19} = 1$
8. $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{3}{9}$; No.

Exercise 3.2.1

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 (Example: (1, 4)).

1. $S =$ _____

Let event A = the sum is even and event B = the first number is prime.

2. $A =$ _____, $B =$ _____
3. $P(A) =$ _____, $P(B) =$ _____
4. $A \cap B =$ _____, $A \cup B =$ _____
5. $P(A \cap B) =$ _____, $P(A \cup B) =$ _____
6. $A' =$ _____, $P(A') =$ _____
7. $P(A) + P(A') =$ _____
8. $P(A|B) =$ _____, $P(B|A) =$ _____; are the probabilities equal?

Example 3.2.2

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 (an outcome is the number of dots that show up).

1. Event $T =$ the outcome is two.
2. Event $A =$ the outcome is an even number.
3. Event $B =$ the outcome is less than four.
4. The complement of A.
5. $A|B$
6. $B|A$
7. $A \cap B$
8. $A \cup B$
9. $A \cup B'$
10. Event $N =$ the outcome is a prime number.
11. Event $I =$ the outcome is seven.

Answer

1. $T = \{2\}$, $P(T) = \frac{1}{6}$
2. $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$
3. $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$
4. $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$
5. $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$
6. $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$
7. $A \cap B = \{2\}$, $P(A \cap B) = \frac{1}{6}$
8. $A \cup B = \{1, 2, 3, 4, 6\}$, $P(A \cup B) = \frac{5}{6}$
9. $A \cup B' = \{2, 4, 5, 6\}$, $P(A \cup B') = \frac{2}{3}$
10. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$
11. A six-sided die does not have seven dots. $P(7) = 0$.

Example 3.2.3

Table 3.1 describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

Table 3.2.1

	Right-handed	Left-handed
Males	43	9
Females	44	4

Let's denote the events $M =$ the subject is male, $F =$ the subject is female, $R =$ the subject is right-handed, $L =$ the subject is left-handed. Compute the following probabilities:

1. $P(M)$
2. $P(F)$
3. $P(R)$
4. $P(L)$
5. $P(M \cap R)$
6. $P(F \cap L)$
7. $P(M \cup F)$
8. $P(M \cup R)$
9. $P(F \cup L)$
10. $P(M')$
11. $P(R|M)$
12. $P(F|L)$
13. $P(L|F)$

Answer

1. $P(M) = 0.52$
2. $P(F) = 0.48$
3. $P(R) = 0.87$
4. $P(L) = 0.13$
5. $P(M \cap R) = 0.43$
6. $P(F \cap L) = 0.04$
7. $P(M \cup F) = 1$
8. $P(M \cup R) = 0.96$
9. $P(F \cup L) = 0.57$
10. $P(M') = 0.48$
11. $P(R|M) = 0.8269$ (rounded to four decimal places)
12. $P(F|L) = 0.3077$ (rounded to four decimal places)
13. $P(L|F) = 0.0833$

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

- **Current page** is licensed CC BY 4.0.
- **3.4: Contingency Tables and Probability Trees** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

3.3: Independent and Mutually Exclusive Events

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events are independent if one of the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \cap B) = P(A)P(B)$

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with** replacement or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether *A* and *B* are independent or dependent, **assume they are dependent until you can show otherwise**.

Example 3.3.1

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), *K* (king) of that suit.

a. Sampling with replacement:
Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the *Q* of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the *Q* of spades again. Your picks are {*Q* of spades, ten of clubs, *Q* of spades}. You have picked the *Q* of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:
Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the *K* of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the *J* of spades. Your picks are {*K* of hearts, three of diamonds, *J* of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice. The probability of picking the three of diamonds is called a conditional probability because it is conditioned on what was picked first. This is true also of the probability of picking the J of spades. The probability of picking the J of spades is actually conditioned on *both* the previous picks.

Exercise 3.3.1

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), *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?

Example 3.3.2

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. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are *QS*, 1*D*, 1*C*, *QD*.
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*, *KH*.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Answer

a. Without replacement; b. With replacement

Exercise 3.3.2

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. *S* = spades, *H* = Hearts, *D* = Diamonds, *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.

a. *QS*, 1*D*, 1*C*, *QD*
b. *KH*, 7*D*, 6*D*, *KH*
c. *QS*, 7*D*, 6*D*, *KS*

Mutually Exclusive **Events**

A and B are **mutually exclusive** events if they cannot occur at the same time. Said another way, if A occurred then B cannot occur and vice versa. This means that A and B do not share any outcomes and $A \cap B = 0$.

For example, suppose the sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let $A = \{1, 2, 3, 4, 5\}$, $B = \{4, 5, 6, 7, 8\}$, and $C = \{7, 9\}$. $A \cap B = \{4, 5\}$. $P(A \cap B) = \frac{2}{10}$ and is not equal to zero. Therefore, A and B are not mutually exclusive. On the other hand, A and C do not have any numbers in common, so $P(A \cap C) = 0$. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example 3.3.3

Flip two fair coins. (This is an experiment.)

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**. (At most one tail means 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 $B = A'$. Also, $P(A) + P(B) = P(A) + P(A') = 1$.
- 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 $B = \{TT\}$, $P(B \cap C) = 0$. B and C are mutually exclusive. (B and C have 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. $F = \{HT, TH, TT\}$. $P(F) = \frac{3}{4}$.

Exercise 3.3.3

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Example 3.3.4

Flip two fair coins. Find the probabilities of the events.

a. Let F = the event of getting at most one tail (zero or one tail).
b. Let G = the event of getting two faces that are the same.
c. Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are F and G mutually exclusive?
e. Let J = the event of getting all tails. Are J and H mutually exclusive?

Answer

Look at the sample space in Example 3.3.3.

a. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(F) = \frac{3}{4}$.
b. Two faces are the same if HH or TT show up. $P(G) = \frac{2}{4}$.
c. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(H) = \frac{2}{4}$.
d. F and G share HH so $P(F \cap G)$ is not equal to zero (0). F and G are not mutually exclusive.
e. Getting all tails occurs when tails shows up on both coins TT. H's outcomes are HH and HT. J and H have nothing in common, so $P(J \cap H) = 0$. J and H are mutually exclusive.

Exercise 3.3.4

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). Find the probability of the following events:

a. Let F = the event of getting the white ball twice.
b. Let G = the event of getting two balls of different colors.
c. Let H = the event of getting white on the first pick.
d. Are F and G mutually exclusive?
e. Are G and H mutually exclusive?

Example 3.3.5

Roll one fair, six-sided die. The sample space is $\{1, 2, 3, 4, 5, 6\}$. Let event A = a face is odd. Then $A = \{1, 3, 5\}$. Let event B = a face is even. Then $B = \{2, 4, 6\}$.

- Find A', the complement of A. The complement of A is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then $C = \{3, 5\}$. Let event D = all even faces smaller than five. Then $D = \{2, 4\}$. $P(C \cap D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. $E = \{1, 2, 3, 4\}$. Are C and E mutually exclusive events? Why or why not?

No. $C = \{3, 5\}$ and $E = \{1, 2, 3, 4\}$. $P(C \cap E) = \frac{1}{6}$. To be mutually exclusive, $P(C \cap E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event $C = \{3, 5\}$ and event $A = \{1, 3, 5\}$. 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, $P(C|A) = \frac{2}{3}$.

Exercise 3.3.5

Let event A = learning Spanish. Let event B = learning German. Then $A \cap B$ = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \cap B) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \cap B) = P(A) * P(B)$

Example 3.3.6

Let event G = taking a math class. Let event H = taking a science class. Then, $G \cap H$ = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \cap H) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \cap H) = P(G) * P(H)$

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.

a. Show that $P(G|H) = P(G)$.

Answer 1

$P(G|H) = \frac{P(G \cap H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$

b. Show $P(G \cap H) = P(G) * P(H)$.

Answer 2

$P(G) * P(H) = (0.6) * (0.5) = 0.3 = P(G \cap H)$.

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent), then knowing that a person is taking a science class would change the chance they are taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

Exercise 3.3.6

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 $P(G \cap O)$?

Example 3.3.7

Let event C= taking an English class. Let event D = taking a speech class. Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \cap D) = 0.225$. Justify your answers to the following questions numerically.

a. Are C and D independent?

b. Are C and D mutually exclusive?

c. What is $P(D|C)$?

Answer

a. Yes, because $P(C|D) = P(C)$.

b. No, because $P(C \cap D)$ is not equal to zero.

c. $P(D|C) = \frac{P(C \cap D)}{P(C)} = \frac{0.225}{0.75} = 0.3$

Exercise 3.3.7

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 $P(B) = 0.4$, $P(D) = 0.3$, and $P(B \cap D) = 0.20$.

a. Find $P(B|D)$.
b. Find $P(D|B)$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Example 3.3.8

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, E = even-numbered card is drawn.

The sample space $S = \{R1, R2, R3, B1, B2, B3, B4, B5\}$. S has eight outcomes.

- $P(R) = \frac{3}{8}$. $P(B) = \frac{5}{8}$. $P(R \cap B) = 0$. (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 $P(R \cap B) = 0$.
- 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.

Exercise 3.3.8

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?

Example 3.3.9

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 be the event that a student is female. Let L be 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:

- $P(F) = 0.6$, $P(L) = 0.5$
- $P(F \cap L) = 0.45$
- $P(L|F) = 0.75$

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.

Check whether $P(F \cap L) = P(F) * P(L)$. We are given that $P(F \cap L) = 0.45$, but $P(F) * P(L) = (0.6) * (0.5) = 0.3$. The events of being female and having long hair are not independent because $P(F \cap L)$ does not equal $P(F) * P(L)$.

Check whether $P(L|F)$ equals $P(L)$. Since we are given that $P(L|F) = 0.75$ and $P(L) = 0.5$, they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent. Knowing that a student is female changes the probability that a student has long hair.

Exercise 3.3.9

Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = 0.44$, $P(F) = 0.56$
- $P(I \cap F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \cup F)$?

Example 3.3.10

a. Toss one fair coin (the coin has two sides, H and T). The outcomes are _____. Count the outcomes. There are ____ outcomes.
b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are _____. Count the outcomes. There are ___ outcomes.
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 to c is the number of outcomes (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 $P(A \cap B)$? If $P(A \cap B) = 0$, then A and B are mutually exclusive.)
h. Are A and B independent? (Hint: Is $P(A \cap B) = P(A) * P(B)$? If $P(A \cap B) = P(A) * P(B)$, then A and B are independent. If not, then they are dependent).

Answer

a. H and T; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. $T1, T2, T3, T4, T5, T6, H1, H2, H3, H4, H5, H6$
e. $A = \{H2, H4, H6\}$; $P(A) = \frac{3}{12}$
f. $B = \{H3\}$; $P(B) = \frac{1}{12}$
g. Yes, because $P(A \cap B) = 0$.
h. $P(A \cap B) = 0$. $P(A) * P(B) = \frac{3}{12}$. $P(A \cap B)$ does not equal $P(A) * P(B)$, so A and B are dependent.

Exercise 3.3.10

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, S the event of picking the white ball in the second drawing.

1. Compute $P(T)$.
2. Compute $P(T|F)$.
3. Are T and F independent?.
4. Are F and S mutually exclusive?
5. Are F and S independent?

This page titled 3.3: Independent and Mutually Exclusive Events is shared under a CC BY license and was authored, remixed, and/or curated by .

3.4: Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a **sample space**, then $P(A \cap B) = P(B) * P(A|B)$. We can think of the intersection symbol as substituting for the word "and".

This rule may also be written as: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

This equation is read as the probability of A given B equals the probability of A and B divided by the probability of B.

If A and B are **independent**, then $P(A|B) = P(A)$. Then $P(A \cap B) = P(A|B) * P(B)$ becomes $P(A \cap B) = P(A) * (B)$ because the $P(A|B) = P(A)$ if A and B are independent.

One easy way to remember the multiplication rule is that the word "and" means that the event has to satisfy two conditions. For example the name drawn from the class roster is to be both a female and a sophomore. It is harder to satisfy two conditions than only one and of course when we multiply fractions the result is always smaller. This reflects the increasing difficulty of satisfying two conditions.

The Addition Rule

If A and B are defined on a sample space, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can think of the union symbol substituting for the word "or". The reason we subtract the intersection of A and B is to keep from double counting elements that are in both A and B.

If A and B are **mutually exclusive**, then $P(A \cap B) = 0$. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes $P(A \cup B) = P(A) + P(B)$.

Example 3.4.1

Klaus is trying to choose where to go on vacation. His two choices are A = New Zealand and B = Alaska.

- Klaus can only afford one vacation. The probability that he chooses A is $P(A) = 0.6$ and the probability that he chooses B is $P(B) = 0.35$.
- $P(A \cap B) = 0$ because Klaus can only afford to take one vacation.
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(A \cup B) = P(A) + P(B) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Example 3.4.2

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(A) = 0.65$. B = the event Carlos is successful on his second attempt. $P(A) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal | that he made the first goal is 0.90.

a. What is the probability that he makes both goals?
b. What is the probability that Carlos makes either the first goal or the second goal?
c. Are A and B independent?
d. Are A and B mutually exclusive?

Answer

a. The problem is asking you to find $P(A \cap B) = P(B \cap A)$. Since $P(B|A) = 0.90$, $P(B \cap A) = P(B|A) * P(A) = (0.90)(0.65) = 0.585$.

   Carlos makes the first and second goals with probability 0.585.

b. The problem is asking you to find $P(A \cup B)$

b. The problem is asking you to find $P(A \cup B)$.
$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.65 + 0.65 - 0.585 = 0.715$

Carlos makes either the first goal or the second goal with probability 0.715.

c. No, they are not, because $P(B \cap A) = 0.585$.

$P(B) * P(A) = (0.65)(0.65) = 0.423$

0.423 ≠ 0.585 = $P(B \cap A)$

So, $P(B \cap A)$ is **not** equal to $P(B) * P(A)$.

d. No, they are not because $P(A \cap B) = 0.585$.

To be mutually exclusive, $P(A \cap B)$ must equal zero.

Exercise 3.4.2

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?

Example 3.4.3

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?

Answer

a. $\frac{28}{150}$
b. $\frac{80}{150}$
c. $\frac{40}{150}$
d. $P(Advanced \cap Intermediate) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.
e. No, these are not independent events.
$P(Novice \cap PracticesFourTimesPerWeek) = 0.0667$
$P(Novice) * P(PracticesFourTimesPerWeek) = 0.0996$
0.0667 ≠ 0.0996

Exercise 3.4.3

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 going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Example 3.4.4

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class | that she enrolls in speech class is 0.25.

Let M = math class, S = speech class, $M|S$ = math class given speech class.

a. What is the probability that Felicity enrolls in math and speech?
   Find $P(M \cap S) = P(M|S) * P(S)$.
b. What is the probability that Felicity enrolls in math or speech classes?
   Find $P(M \cup S) = P(M) + P(S) - P(M \cap S)$.
c. Are M and S independent? Is $P(M|S) = P(M)$?
d. Are M and S mutually exclusive? Is $P(M \cap S) = 0$?

Answer

a. 0.1625, b. 0.6875, c. No, d. No

Exercise 3.4.4

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 $P(B \cap D)$.
b. Find $P(B \cup D)$.

Example 3.4.5

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 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?

Answer

a. $P(B) = 0.143$; $P(N) = 0.85$

b. $P(N|B) = 0.02$

c. $P(B \cap N) = P(B) * P(N|B) = (0.143)(0.02) = 0.0029$

d. $P(B \cup N) = P(B) + P(N) - P(B \cap N) = 0.143 + 0.85 - 0.0029 = 0.9901$

e. No. $P(N) = 0.85$; $P(N|B) = 0.02$. So, $P(N|B)$ does not equal $P(N)$.

f. No. $P(B \cap N) = 0.0029$. For B and N to be mutually exclusive, $P(B \cap N)$ must be zero.

Exercise 3.4.5

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 going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Example 3.4.6

Refer to the information in Example 3.4.5. 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 $P(B \cap P) = P(B) * P(P|B)$.
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 $P(P) = 1 - P(N)$.

Answer

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Exercise 3.4.6

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 $P(D|B) = 0.5$.

1. Find $P(B')$.
2. Find $P(D \cap B)$.
3. Find $P(B|D)$.
4. Find $P(D \cap B')$.
5. Find $P(D|B')$.

This page titled 3.4: Two Basic Rules of Probability is shared under a CC BY license and was authored, remixed, and/or curated by .

3.5: Contingency Tables and Probability Trees

Contingency Tables

A **contingency table** provides a way of portraying data 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, we will use contingency tables again, but in another manner.

Example 3.5.1

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

Table 3.5.1

	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

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

Calculate the following probabilities using the table.

a. Find P(Driver is a cell phone user).

Answer

a. $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$

b. Find P(Driver had no violation in the last year).

Answer

b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$

c. Find P(Driver had no violation in the last year $\cap$ was a cell phone user).

Answer

c. $\frac{280}{755}$

d. Find P(Driver is a cell phone user $\cup$ driver had no violation in the last year).

Answer

d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$

e. Find P(Driver is a cell phone user | driver had a violation in the last year).

Answer

e. $\frac{25}{70}$ (The sample space is reduced to the number of drivers who had a violation.)

f. Find P(Driver had no violation last year | driver was not a cell phone user)

Answer

f. $\frac{405}{450}$ (The sample space is reduced to the number of drivers who were not cell phone users.)

Exercise 3.5.1

Table 3.5.2 shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

Table 3.5.2

1. What is P(athlete stretches before exercising)?
2. What is P(athlete stretches before exercising|no injury in the last year)?

Example 3.5.2

Table 3.5.3 shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	The coastline	Near lakes and streams	On mountain peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Table 3.5.3 Hiking Area Preference

a. Complete the table.

Answer

a.

Table 3.5.4 Hiking Area Preference

Sex	The coastline	Near lakes and streams	On mountain peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

b. Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \cap C)$.
2. 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.

Answer

b.

1. $P(F \cap C) = \frac{18}{100}$ = 0.18
2. P(F)P(C) = $(\frac{45}{100})(\frac{34}{100})$ = (0.45)(0.34) = 0.153

$P(F \cap C) \neq P$(F)P(C), so the events 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 = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: P(___|___) = ___.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

Answer

c.

1.The word 'given' tells you that this is a conditional.

2.P(M|L) = $\frac{25}{41}$

3.No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = prefers mountain peaks.

1. Find P(F).
2. Find P(P).
3. Find $P(F \cap P)$.
4. Find $P(F \cup P)$.

Answer

d.

1. P(F) = $\frac{45}{100}$
2. P(P) = $\frac{25}{100}$
3. $P(F \cap P) = \frac{11}{100}$
4. $P(F \cup P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Exercise 3.5.2

Table 3.5.5 shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Table 3.5.5

Gender	Lake path	Hilly path	Wooded path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

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?

Example 3.5.3

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 1/15 and the probability he is not caught is 4/15. If he goes out the second door, the probability he gets caught by Alissa is 1/12 and the probability he is not caught is 3/12. The probability that Alissa catches Muddy coming out of the third door is 1/6 and the probability she does not catch Muddy is 1/6. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is 1/3.

Table 3.5.6 Door Choice

Caught or not	Door one	Door two	Door three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is $P(DoorOne \cap Caught)$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is $P(DoorOne \cap NotCaught)$

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.

Answer

a.

Table 3.5.7 Door Choice

Caught or not	Door one	Door two	Door three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. What is the probability that Alissa does not catch Muddy?

Answer

b. $\frac{41}{60}$

c. What is the probability that Muddy chooses Door One ∩ Door Two given that Muddy is caught by Alissa?

Answer

c. $\frac{9}{19}$

Example 3.5.4

Table 3.5.8 contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Table 3.5.8 United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

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					

TOTAL each column and each row. Total data = 4,520.7

1. Find $P(2009 \cap Robbery)$.
2. Find $P(2010 \cap Burglary)$.
3. Find $P(2010 \cup Burglary)$.
4. Find *P*(2011|Rape).
5. Find *P*(Vehicle|2008).

Answer

1. 0.0294
2. 0.1551
3. 0.7165
4. 0.2365
5. 0.2575

Exercise 3.5.3

Table 3.5.9 relates the weights and heights of a group of individuals participating in an observational study.

Table 3.5.9

Weight/height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

1. Find the total for each row and column
2. Find the probability that a randomly chosen individual from this group is Tall.
3. Find the probability that a randomly chosen individual from this group is Obese and Tall.
4. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
5. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
6. Find the probability a randomly chosen individual from this group is Tall and Underweight.
7. Are the events Obese and Tall independent?

Tree Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams can be used to visualize and solve conditional probabilities.

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 3.5.5

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.

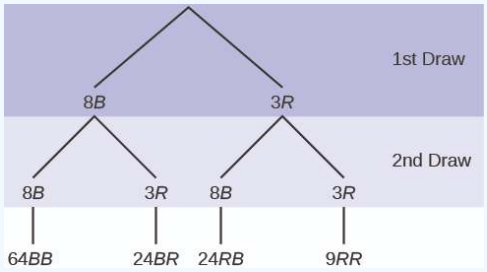


Figure 3.5.1 Total = 64 + 24 + 24 + 9 = 121

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:

R1R1; R1R2; R1R3; R2R1; R2R2; R2R3; R3R1; R3R2; R3R3

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, ...

Answer

a. B1R1; B1R2; B1R3; B2R1; B2R2; B2R3; B3R1; B3R2; B3R3; B4R1; B4R2; B4R3; B5R1; B5R2; B5R3; B6R1; B6R2; B6R3; B7R1; B7R2; B7R3; B8R1; B8R2; B8R3

b. Using the tree diagram, calculate *P*(RR).

Answer

b. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. Using the tree diagram, calculate *P*(RB ∪ BR).

Answer

c. $P(RB \cup BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. Using the tree diagram, calculate *P*(R on 1st draw ∩ B on 2nd draw).

Answer

d. $P(\text{R on 1st draw} \cap \text{B on 2nd draw}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. Using the tree diagram, calculate *P*(R on 2nd draw|B on 1st draw).

Answer

e. P(R on 2nd draw|B on 1st draw) = P(R on 2nd|B on 1st) = $\frac{24}{88} = \frac{3}{11}$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $\frac{24}{88} = \frac{3}{11}$.

f. Using the tree diagram, calculate P(BB).

Answer

f. P(BB) = $\frac{64}{121}$

g. Using the tree diagram, calculate P(B on the 2nd draw|R on the first draw).

Answer

g. P(B on 2nd draw|R on 1st draw) = $\frac{8}{11}$

There are 9 + 24 outcomes that have R on the first draw (9 RR and 24 RB). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Exercise 3.5.4

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).

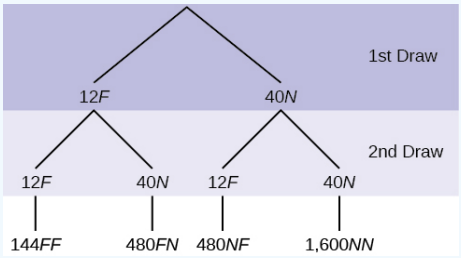


Figure 3.5.2

Example 3.5.6

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second marble. Following is 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, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

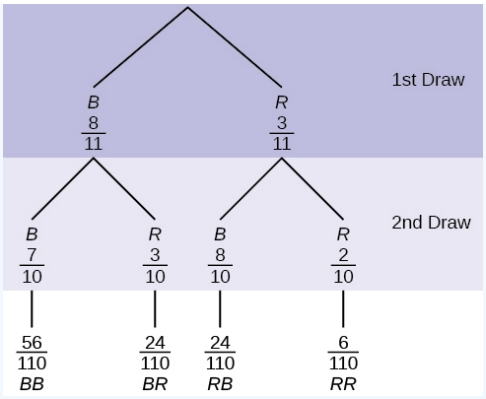


Figure 3.5.3

Total = $\frac{56+24+24+6}{110} = \frac{110}{110} = 1$

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**, so that on the second draw there are ten marbles left in the urn.

Calculate the following probabilities using the tree diagram.

a. P(RR) = _____

Answer

a. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

b. Fill in the blanks:

$P(RB \cup BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right)$ + (___)(___) = $\frac{48}{110}$

Answer

b. $P(RB \cup BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

c. P(R on 2nd|B on 1st) =

Answer

c. P(R on 2nd|B on 1st) = $\frac{3}{10}$

d. Fill in the blanks.

P(R on 1st ∩ B on 2nd) = (___)(___) = $\frac{24}{100}$

Answer

d. $P(R \text{ on 1st } \cap B \text{ on 2nd}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{110}$

e. Find P(BB).

Answer

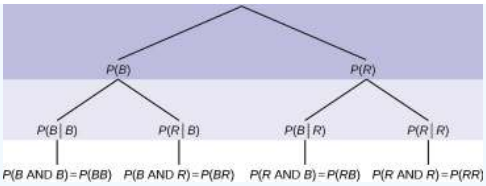
e. $P(\text{BB}) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

f. Find P(B on 2nd|R on 1st).

Answer

f. Using the tree diagram, P(B on 2nd|R on 1st) = P(R|B) = $\frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.



- P(R|R) here means P(R on 2nd|R on 1st)
- P(B|R) here means P(B on 2nd|R on 1st)
- P(R|B) here means P(R on 2nd|B on 1st)
- P(B|B) here means P(B on 2nd|B on 1st)

Exercise 3.5.5

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 is labeled with all possible probabilities.

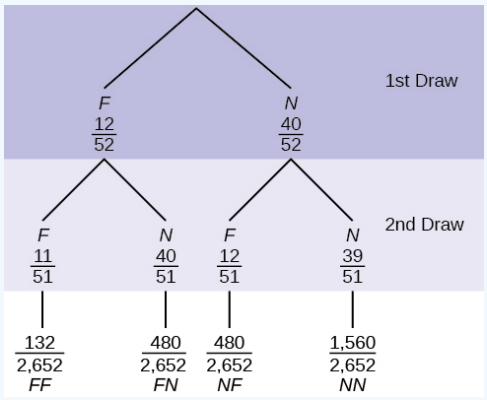
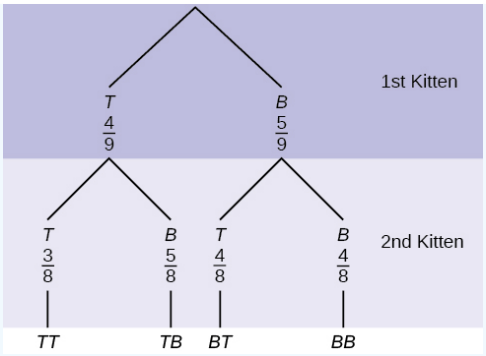


Figure 3.5.4

1. Find $P(FN \cup NF)$.
2. Find P(N|F).
3. Find P(at most one face card).
   Hint: "At most one face card" means zero or one face card.
4. Find P(at least on face card).
   Hint: "At least one face card" means one or two face cards.

Example 3.5.7

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.



1. What is the probability that both kittens are tabby?

   $a \cdot \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\ b \cdot \left(\frac{4}{9}\right)\left(\frac{4}{9}\right)\ c \cdot \left(\frac{4}{9}\right)\left(\frac{3}{8}\right)\ d \cdot \left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$
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?

Answer

1. c, 2. d, 3. $\frac{4}{8}$, 4. $\frac{32}{72}$

Exercise 3.5.6

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?

This page titled 3.5: Contingency Tables and Probability Trees is shared under a CC BY license and was authored, remixed, and/or curated by .

3.6: Chapter 3 Key Terms

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

Contingency Table

the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

Dependent Events

If two events are NOT independent, then we say that they are dependent.

Equally Likely

Each outcome of an experiment has the same probability.

Event

a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a sample space and is usually denoted by *S*. An event is an arbitrary subset in *S*. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital, italicized letters such as *A*, *B*, *C*, and so on.

Experiment

a planned activity carried out under controlled conditions

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A \cap B) = 0$.

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let *S* denote the sample space and *A* and *B* are two events in *S*. Then:

- $0 \leq P(A) \leq 1$
- If *A* and *B* are any two mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$.
- $P(S) = 1$

Sample Space

the set of all possible outcomes of an experiment

Sampling with Replacement

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement

When sampling is done without replacement, each member of a population may be chosen only once.

The Complement Event

The complement of event *A* consists of all outcomes that are NOT in *A*.

The Conditional Probability of $A|B$

P(A|B) is the probability that event *A* will occur given that the event *B* has already occurred.

The Intersection: the $\cap$ Event

An outcome is in the event $A \cap B$ if the outcome is in both $A \cap B$ at the same time.

The Union: the $\cup$ Event

An outcome is in the event $A \cup B$ if the outcome is in *A* or is in *B* or is in both *A* and *B*.

Tree Diagram

the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies

---

3.6: Chapter 3 Key Terms is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.7: Chapter 3 Review

3.1 Terminology

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

3.2 Independent and Mutually Exclusive Events

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are **dependent**.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent.

Events A and B are **mutually exclusive** events when they do not have any outcomes in common.

3.3 Two Basic Rules of Probability

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. See the Formula Review section for details.

3.4 Contingency Tables and Probability Trees

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.

Tree diagrams use branches to show the different outcomes of experiments and make complex probability questions easy to visualize.

3.7: Chapter 3 Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.8: Chapter 3 Formula Review

3.1 Terminology

A and *B* are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

3.2 Independent and Mutually Exclusive Events

If A and B are independent, $P(A \cap B) = P(A)P(B), P(A|B) = P(A)$ and $P(B|A) = P(B)$

If A and B are mutually exclusive, $P(A \cup B) = P(A) + P(B)$ and $P(A \cap B) = 0$

3.3 Two Basic Rules of Probability

The multiplication rule: $P(A \cap B) = P(A|B)P(B)$

The addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

3.8: Chapter 3 Formula Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.9: Chapter 3 Homework

3.2 Terminology

1. 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 for parts 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 be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be 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.

Use the following information to answer the next four exercises. 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.

2. Find $P(H)$.

3. Find $P(N)$.

4. Find $P(F)$.

5. Find $P(C)$.

Use the following information to answer the next six exercises. 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.

6. Find P(B).

7. Find P(G).

8. Find P(P).

9. Find P(R).

10. Find P(Y).

11. Find P(O).

Use the following information to answer the next six exercises. 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.

12. Find $P(A)$.

13. Find $P(E)$.

14. Find $P(F)$.

15. Find $P(N)$.

16. Find $P(O)$.

17. Find $P(S)$.

18. What is the probability of drawing a red card in a standard deck of 52 cards?

19. What is the probability of drawing a club in a standard deck of 52 cards?

20. What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

21. What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Use the following information to answer the next two exercises. 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.

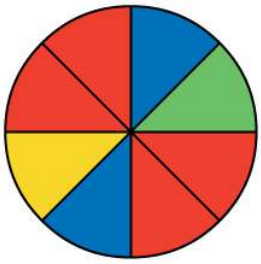


Figure 3.9.1

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.

22. If you land on Y, you get the biggest prize. Find $P(Y)$.

23. If you land on red, you don't get a prize. What is $P(R)$?

Use the following information to answer the next ten exercises. 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.

24. Write the symbols for the probability that a player is not an outfielder.

25. Write the symbols for the probability that a player is an outfielder or is a great hitter.

26. Write the symbols for the probability that a player is an infielder and is not a great hitter.

27. Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

28. Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

29. Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

30. Write the symbols for the probability that of all the great hitters, a player is an outfielder.

31. Write the symbols for the probability that a player is an infielder or is not a great hitter.

32. Write the symbols for the probability that a player is an outfielder and is a great hitter.

33. Write the symbols for the probability that a player is an infielder.

34. What is the word for the set of all possible outcomes?

35. What is conditional probability?

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
What is the sample space?

37. What is the sum of the probabilities of an event and its complement?

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

38. What does $P(E|M)$ mean in words?

39. What does $P(E \cup M)$ mean in words?

3.3 Independent and Mutually Exclusive Events

40. E and F are mutually exclusive events. $P(E) = 0.4$ and $P(F) = 0.5$. Find $P(E|F)$.

41. J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

42. U and V are mutually exclusive events. $P(U) = 0.26$ and $P(V)=0.37$. Find the following probabilities:

1. $P(\text{U} \cap \text{V}) =$
2. $P(\text{U}|\text{V}) =$
3. $P(\text{U} \cup \text{V}) =$

43. Q and R are independent events. $P(Q)=0.4$ and $P(Q \cap R)=0.1$. Find $P(R)$.

3.4 Two Basic Rules of Probability

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians 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.

44. Find $P(C)$.

45. Find $P(L)$.

46. Find $P(C|L)$.

47. In words, what is $C|L$?

48. Find $P(L \cap C)$.

49. In words, what is $L \cap C$?

50. Are L and C independent events? Show why or why not.

51. Find $P(L \cup C)$.

52. In words, what is $L \cup C$?

53. Are L and C mutually exclusive events? Show why or why not.

3.5 Contingency Tables and Probability Trees

Use the following information to answer the next four exercises. Table 3.9.1 shows a random sample of musicians and how they learned to play their instruments.

Table 3.9.1

Gender	Self-taught	Studied in school	Private instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

54. Find *P*(musician is a female).

55. Find *P*(musician is a male ∩ had private instruction).

56. Find *P*(musician is a female ∪ is self taught).

57. Are the events “being a female musician” and “learning music in school” mutually exclusive events?

58. 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. Let *C* = a man develops cancer in his lifetime and *P* = man has at least one false positive. Construct a tree diagram of the situation.

Use the following information to answer the next seven exercises. 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.

59. Complete the table using the data provided above. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Table 3.9.2 Smoking Levels by Ethnicity

Smoking level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

60. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

61. Find the probability that the person was Latino.

62. 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." Also, find the probability.

63. In words, explain what it means to pick one person from the study who is "Japanese American ∪ smokes 21 to 30 cigarettes per day." Also, find the probability.

64. In words, explain what it means to pick one person from the study who is "Japanese American | that person smokes 21 to 30 cigarettes per day." Also, find the probability.

65. Prove that smoking level/day and ethnicity are dependent events.

Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

66. Suppose that you randomly draw two cards, one at a time, **with replacement**.
Let G_1 = first card is green
Let G_2 = second card is green

1. Draw a tree diagram of the situation.
2. Find $P(G_1 \cap G_2)$.
3. Find *P*(at least one green).
4. Find $P(G_2|G_1)$.
5. Are G_2 and G_1 independent events? Explain why or why not.

67. Suppose that you randomly draw two cards, one at a time, **without replacement**.
G_1 = first card is green
G_2 = second card is green

1. Draw a tree diagram of the situation.
2. Find $P(G_1 \cap G_2)$.
3. Find *P*(at least one green).
4. Find $P(G_2|G_1)$.
5. Are G_2 and G_1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. 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; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

68. Complete the following.

1. Construct a table or a tree diagram of the situation.
2. Find *P*(driver is female).

3. Find P(driver is age 65 or over | driver is female).
4. Find P(driver is age 65 or over ∩ female).
5. In words, explain the difference between the probabilities in part c and part d.
6. Find P(driver is age 65 or over).
7. Are being age 65 or over and being female mutually exclusive events? How do you know?

69. Suppose that 10,000 U.S. 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 out of the age 20–64 group, a randomly selected driver is female.

70. 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.

1. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
2. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
3. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
4. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

71. 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 out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

1. Based on the given data, find $P(H)$ and $P(T)$.
2. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
3. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
4. Use the tree to find the probability of obtaining at least one head.

3.2 Terminology

72.

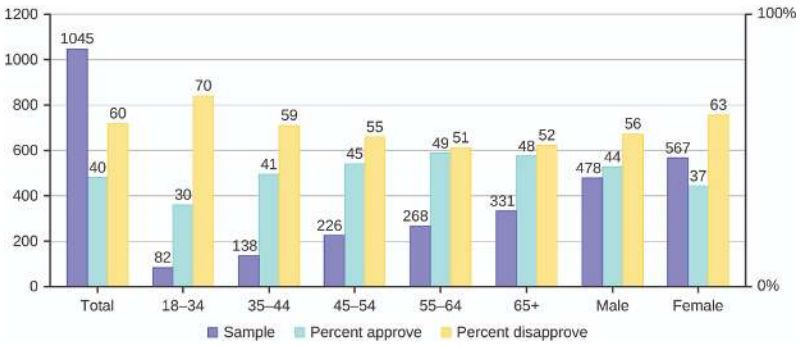


Figure 3.9.2

The graph in Figure 3.9.2 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.

1. Define three events in the graph.
2. Describe in words what the entry 40 means.
3. Describe in words the complement of the entry in question 2.
4. Describe in words what the entry 30 means.
5. Out of the males and females, what percent are males?
6. Out of the females, what percent disapprove of Mayor Ford?
7. Out of all the age groups, what percent approve of Mayor Ford?
8. Find P(Approve|Male).
9. Out of the age groups, what percent are more than 44 years old?

10. Find $P(\text{Approve}|\text{Age} < 35)$.

73. 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.

3.3 Independent and Mutually Exclusive Events

Use the following information to answer the next 12 exercises. 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.

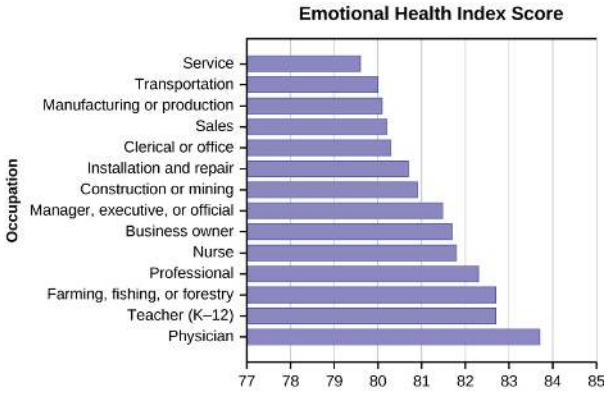


Figure 3.9.3

74. Find the probability that an Emotional Health Index Score is 82.7.

75. Find the probability that an Emotional Health Index Score is 81.0.

76. Find the probability that an Emotional Health Index Score is more than 81?

77. Find the probability that an Emotional Health Index Score is between 80.5 and 82?

78. If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

79. What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

80. What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

81. What occupation has the highest emotional index score?

82. What occupation has the lowest emotional index score?

83. What is the range of the data?

84. Compute the average EHIS.

85. 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?

3.4 Two Basic Rules of Probability

86. On February 28, 2013, a Field Poll Survey reported that 61% of California 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 olds (California registered voters), the approval rating was 78%. Six in ten California 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.

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.

1. Find $P(C)$.
2. Find $P(B)$.
3. Find $P(C|A)$.
4. Find $P(B|C)$.
5. In words, what is $C|A$?
6. In words, what is $B|C$?
7. Find $P(C \cap B)$.
8. In words, what is $C \cap B$?
9. Find $P(C \cup B)$.
10. Are C and B mutually exclusive events? Show why or why not.

87. After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

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?

Use the following information to answer the next three exercises. 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 of 38 numbers, and each number is assigned to a color and a range.

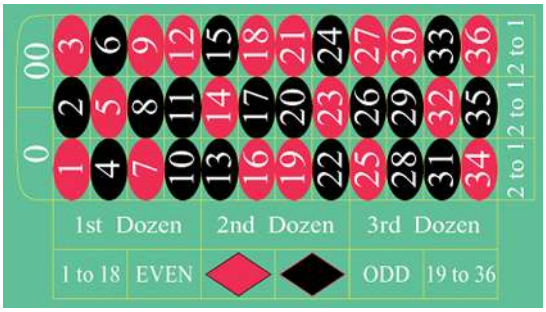


Figure 3.9.4 (credit: film8ker/wikibooks)

88.

1. List the sample space of the 38 possible outcomes in roulette.
2. You bet on red. Find *P*(red).
3. You bet on -1st 12- (1st Dozen). Find *P*(-1st 12-).
4. You bet on an even number. Find P(even number).
5. Is getting an odd number the complement of getting an even number? Why?
6. Find two mutually exclusive events.
7. Are the events Even and 1st Dozen independent?

89. 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

90. Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

91. 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

1. List the sample space.
2. $P(G) =$ _____
3. $P(G|E) =$ _____
4. $P(G \cap E) =$ _____
5. $P(G \cup E) =$ _____
6. Are G and E mutually exclusive? Justify your answer numerically.

92. Roll two fair dice separately. Each die has six faces.

1. List the sample space.
2. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
3. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.
4. In words, explain what "$P(A|B)$" represents. Find $P(A|B)$.
5. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
6. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

93. 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.

1. List the sample space.
2. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(A)$.
3. 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.
4. 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.

94. An experiment consists of first rolling a die and then tossing a coin.

1. List the sample space.
2. 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).
3. 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.

95. An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

1. List the sample space.
2. Let *A* be the event that there are at least two tails. Find *P*(*A*).
3. 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.

96. Consider the following scenario:
Let $P(C) = 0.4$.
Let $P(D) = 0.5$.
Let $P(C|D) = 0.6$.

1. Find $P(C \cap D)$.
2. Are *C* and *D* mutually exclusive? Why or why not?
3. Are *C* and *D* independent events? Why or why not?
4. Find $P(C \cup D)$.
5. Find $P(D|C)$.

97. *Y* and *Z* are independent events.

1. Rewrite the basic Addition Rule $P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z)$ using the information that *Y* and *Z* are independent events.
2. Use the rewritten rule to find $P(Z)$ if $P(Y \cup Z) = 0.71$ and $P(Y) = 0.42$.

98. *G* and *H* are mutually exclusive events. $P(G) = 0.5 P(H) = 0.3$

1. Explain why the following statement MUST be false: $P(H|G) = 0.4$.
2. Find $P(H \cup G)$.
3. Are *G* and *H* independent or dependent events? Explain in a complete sentence.

99. Approximately 281,000,000 people over age 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.

Let *E* = speaks English at home, *E*′ = speaks another language at home, and *S* = speaks Spanish.

Finish each probability statement by matching the correct answer.

Table 3.9.3

Probability Statements	Answers
a. $P(E') =$	i. 0.8043
b. $P(E) =$	ii. 0.623
c. $P(S \cap E') =$	iii. 0.1957
d. $P(S|E') =$	iv. 0.1219

100. In 1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let *G* = won green card.

1. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
2. 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.
3. Are *G* and *F* independent or dependent events? Justify your answer numerically and also explain why.
4. Are *G* and *F* mutually exclusive events? Justify your answer numerically and explain why.

101. 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. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let R = money returned, E = economics classes, and O = other classes

1. Write a probability statement for the overall percent of money returned.
2. Write a probability statement for the percent of money returned out of the economics classes.
3. Write a probability statement for the percent of money returned out of the other classes.
4. Is money being returned independent of the class? Justify your answer numerically and explain it.
5. 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.

102. The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Table 3.9.4

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

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

1. Yes, because P(hit by Hank Aaron|hit is a double) = P(hit by Hank Aaron)
2. No, because P(hit by Hank Aaron|hit is a double) ≠ P(hit is a double)
3. No, because P(hit is by Hank Aaron|hit is a double) ≠ P(hit by Hank Aaron)
4. Yes, because P(hit is by Hank Aaron|hit is a double) = P(hit is a double)

103. United Blood Services 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 bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh- factor, while 52% of people have type O or Rh- factor.

1. Find the probability that a person has both type O blood and the Rh- factor.
2. Find the probability that a person does NOT have both type O blood and the Rh- factor.

104. 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.

1. Find the probability that a course has a final exam or a research project.
2. Find the probability that a course has NEITHER of these two requirements.

105. 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 (he can't eat it).
b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

106. 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 = event that a student takes a distance learning class and E = event that a student is a part time student

1. Find $P(D \cap E)$.
2. Find $P(E|D)$.
3. Find $P(D \cup E)$.
4. Using an appropriate test, show whether D and E are independent.

5. Using an appropriate test, show whether D and E are mutually exclusive.

3.5 Contingency Tables and Probability Trees

Use the information in the Table 3.9.5 to answer the next eight exercises. 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.

Table 3.9.5

Up for reelection:	Democratic party	Republican party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

107. What is the probability that a randomly selected senator has an "Other" affiliation?

108. What is the probability that a randomly selected senator is up for reelection in November 2016?

109. What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

110. What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

111. 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?

112. 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?

113. The events "Republican" and "Up for reelection in 2016" are _____

1. mutually exclusive.
2. independent.
3. both mutually exclusive and independent.
4. neither mutually exclusive nor independent.

114. The events "Other" and "Up for reelection in November 2016" are _____

1. mutually exclusive.
2. independent.
3. both mutually exclusive and independent.
4. neither mutually exclusive nor independent.

115. Table 3.9.6 gives the number of participants in the recent National Health Interview Survey who had been treated for cancer in the previous 12 months. The results are sorted 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.

Table 3.9.6

Race and sex	15–24	25–40	41–65	Over 65	TOTALS
White, male	1,165	2,036	3,703		8,395
White, female	1,076	2,242	4,060		9,129
Black, male	142	194	384		824
Black, female	131	290	486		1,061
All others					
TOTALS	2,792	5,279	9,354		21,081

Do not include "all others" for parts f and g.

a. Fill in the column for cancer treatment for individuals over age 65.
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 male.
g. Out of the individuals over age 65, find the probability that a randomly selected individual was a black or white male.

Use the following information to answer the next two exercises. The table of data obtained from *www.baseball-almanac.com* shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

Table 3.9.7

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

116. Find *P*(hit was made by Babe Ruth).

1. $\frac{1518}{2873}$
2. $\frac{2873}{12351}$
3. $\frac{583}{12351}$
4. $\frac{4189}{12351}$

117. Find *P*(hit was made by Ty Cobb|The hit was a Home Run).

1. $\frac{4189}{12351}$
2. $\frac{114}{1720}$
3. $\frac{1720}{4189}$
4. $\frac{114}{12351}$

118. Table 3.9.8 identifies a group of children by one of four hair colors, and by type of hair.

Table 3.9.8

Hair type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

1. Complete the table.
2. What is the probability that a randomly selected child will have wavy hair?
3. What is the probability that a randomly selected child will have either brown or blond hair?
4. What is the probability that a randomly selected child will have wavy brown hair?
5. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
6. If B is the event of a child having brown hair, find the probability of the complement of B.
7. In words, what does the complement of B represent?

119. 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.

Table 3.9.9

Shirt #	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

1. Find the probability that his shirt number is from 1 to 33.
2. Find the probability that he weighs at most 210 pounds.
3. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
4. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
5. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

Use the following information to answer the next two exercises. 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, $P(\text{H}) = \frac{2}{3}$ and $P(\text{T}) = \frac{1}{3}$ where H is heads and T is tails.

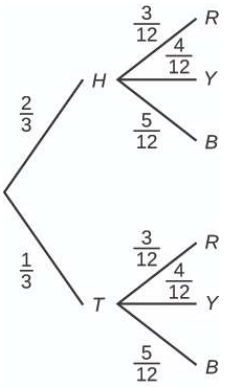


Figure 3.9.5

120. Find P(tossing a Head on the coin AND a Red bead)

1. $\frac{2}{3}$
2. $\frac{5}{15}$
3. $\frac{6}{36}$
4. $\frac{5}{36}$

121. Find P(Blue bead).

1. $\frac{15}{36}$
2. $\frac{10}{36}$
3. $\frac{10}{12}$
4. $\frac{6}{36}$

122. 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. (How many cookies did he take?)

1. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
2. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
3. For each complete path through the tree, write the event it represents and find the probabilities.
4. Let S be the event that both cookies selected were the same flavor. Find $P(S)$.

5. Let T be the event that the cookies selected were different flavors. Find $P(T)$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
6. Let U be the event that the second cookie selected is a butter cookie. Find $P(U)$.

3.9: Chapter 3 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.10: Chapter 3 Solutions

1.

1. $P(L') = P(S)$
2. $P(M \cup S)$
3. $P(F \cap L)$
4. $P(M|L)$
5. $P(L|M)$
6. $P(S|F)$
7. $P(F|L)$
8. $P(F \cup L)$
9. $P(M \cap S)$
10. $P(F)$

3. $P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

5. $P(C) = \frac{5}{42} = 0.12$

7. $P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

9. $P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

11. $P(O) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

13. $P(E) = \frac{47}{194} = 0.24$

15. $P(N) = \frac{23}{194} = 0.12$

17. $P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

19. $\frac{13}{52} = \frac{1}{4} = 0.25$

21. $\frac{3}{6} = \frac{1}{2} = 0.5$

23. $P(R) = \frac{4}{8} = 0.5$

25. $P(O \cup H)$

27. $P(H|I)$

29. $P(N|O)$

31. $P(I \cup N)$

33. $P(I)$

35. The likelihood that an event will occur given that another event has already occurred.

37. 1

39. the probability of landing on an even number or a multiple of three

41. $P(J) = 0.3$

43. $P(Q \cap R) = P(Q)P(R)$

$0.1 = (0.4)P(R)$

$P(R) = 0.25$

45. 0.376

47. $C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

49. $L \cap C$ is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

51. 0.6492

53. No, because $P(L \cap C)$ does not equal 0.

55. $P(\text{musician is a male} \cap \text{had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$.

57. The events are not mutually exclusive. It is possible to be a female musician who learned music in school.

58.

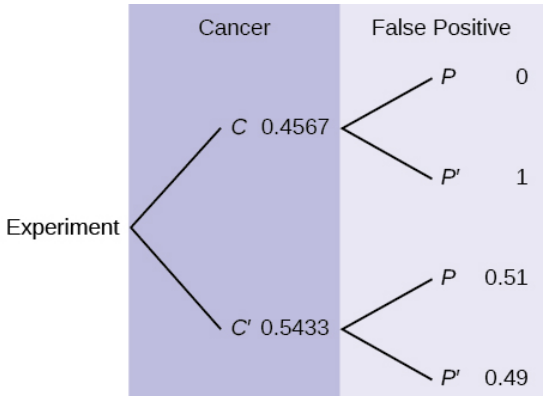


Figure 3.10.1

60. $\frac{35,065}{100,450}$

62. To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

64. To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4715}{15,273}$.

66.

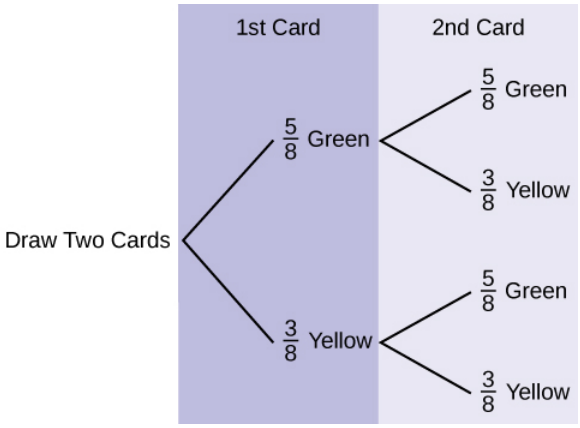


Figure 3.10.2

1. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
2. $P(\text{at least one green}) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
3. $P(G|G) = \frac{5}{8}$
4. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

68.

	<20>	20–64	>64	Totals
Female	" class="lt-biz-79015">0.0244	0.3954	64" class="lt-biz-79015">0.0661	0.486
Male	" class="lt-biz-79015">0.0259	0.4186	64" class="lt-biz-79015">0.0695	0.514
Totals	" class="lt-biz-79015">0.0503	0.8140	64" class="lt-biz-79015">0.1356	1

Table 3.10.1

1. $P(F) = 0.486$
2. $P(> 64|F) = 0.1361$
3. $P(> 64 \text{ and } F) = P(F) * P(> 64|F) = (0.486) * (0.1361) = 0.0661$
4. $P(> 64|F)$ is the percentage of female drivers who are 65 or older and $P(> 64 \cap F)$ is the percentage of drivers who are female and 65 or older.
5. $P(> 64) = P(> 64 \cap F) + P(> 64 \cap M) = 0.1356$
6. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(> 64 \cap F) = 0.0661$.

70.

	Car, truck or van	Walk	Public transportation	Other	Totals
Alone	0.7318				
Not alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

Table 3.10.2

1. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$.
2. Make the same assumptions as in (b) we have: $(0.7708)(1,000) = 771$
3. $(0.1332)(1,000) = 133$

73.

1. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given. The probabilities should be multiplied, not added, and probability is never greater than 100%.
2. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

75. 0

77. 0.3571

79. 0.2142

81. Physician (83.7)

83. $83.7 - 79.6 = 4.1$

85. $P(\text{Occupation} < 81.3) = 0.5$

87.

1. The Forum Research surveyed 1,046 Torontonians.
2. 58%

3. 42% of 1,046 = 439 (rounding to the nearest integer)
4. 0.57
5. 0.60.

89.

1. $P(\text{Betting on two line that touch each other on the table}) = \frac{6}{38}$.
2. $P(\text{Betting on three numbers in a line}) = \frac{3}{38}$
3. $P(\text{Betting on one number}) = \frac{1}{38}$
4. $P(\text{Betting on four number that touch each other to form a square}) = \frac{4}{38}$.
5. $P(\text{Betting on two number that touch each other on the table}) = \frac{2}{38}$
6. $P(\text{Betting on } 0-00-1-2-3) = \frac{5}{38}$
7. $P(\text{Betting on } 0-1-2;\ \text{or } 0-00-2;\ \text{or } 00-2-3) = \frac{3}{38}$

91.

1. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
2. $\frac{5}{8}$
3. $\frac{2}{3}$
4. $\frac{2}{8}$
5. $\frac{6}{8}$
6. No, because $P(G \cap E)$ does not equal 0.

93. Note: The coin toss is independent of the card picked first.

1. $\{(G,H)(G,T)(B,H)(B,T)(R,H)(R,T)\}$
2. $P(A) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
3. Yes, *A* and *B* are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \cap B) = 0$
4. No, *A* and *C* are not mutually exclusive because they can occur at the same time. In fact, *C* includes all of the outcomes of *A*. If the card chosen is blue, it is also (red or blue). $P(A \cap C) = P(A) = \frac{3}{20}$.

95.

1. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$
2. $\frac{4}{8}$
3. Yes, because if *A* has occurred, it is impossible to obtain two tails. In other words, $P(A \cap B) = 0$.

97.

1. If *Y* and *Z* are independent, then $P(Y \cap Z) = P(Y)P(Z)$, so $P(Y \cup Z) = P(Y) + P(Z) - P(Y)P(Z)$.
2. 0.5

99. iii; i; iv; ii

101.

1. $P(R) = 0.44$
2. $P(R|E) = 0.56$
3. $P(R|O) = 0.31$
4. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate, $P(R|E) \neq P(R)$.
5. No, this study definitely does not support that notion; in fact, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively, $P(R|E) > P(R)$.

103.

1. $P(\text{ type O} \cup \text{Rh}-) = P(\text{ type O}) + P(\text{Rh}-) - P(\text{ type O} \cap \text{Rh}-)$

$0.52 = 0.43 + 0.15 - P(\text{ type } O \cap \text{Rh}-)$; solve to find $P(\text{ type O} \cap \text{Rh}-) = 0.06$

6% of people have type O, Rh- blood

2. $P(\text{ NOT (type O } \cap \text{Rh}-)) = 1 - P(\text{ type O} \cap \text{Rh}-) = 1 - 0.06 = 0.94$

94% of people do not have type O, Rh- blood

105.

1. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
2. $P(C \cup N) = P(C) + P(N) - P(C \cap N) = 0.36 + 0.12 - 0.08 = 0.40$
3. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(C \cup N) = 1 - 0.40 = 0.60$

107. 0

109. $\frac{10}{67}$

111. $\frac{10}{34}$

113. d

115.

Race and sex	1–14	15–24	25–64	Over 64	TOTALS
White, male	210	3,360	13,610	4,870	22,050
White, female	80	580	3,380	890	4,930
Black, male	10	460	1,060	140	1,670
Black, female	0	40	270	20	330
All others				100	
TOTALS	310	4,650	18,780	6,020	29,760

Table 3.10.3

Race and sex	1–14	15–24	25–64	Over 64	TOTALS
White, male	210	3,360	13,610	4,870	22,050
White, female	80	580	3,380	890	4,930
Black, male	10	460	1,060	140	1,670
Black, female	0	40	270	20	330
All others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

Table 3.10.4

1. $\frac{22,050}{29,760}$
2. $\frac{330}{29,760}$
3. $\frac{2,000}{29,760}$
4. $\frac{23,720}{29,760}$
5. $\frac{5,010}{6,020}$

117. b

119.

1. $\frac{26}{106}$
2. $\frac{33}{106}$
3. $\frac{21}{106}$
4. $\left(\frac{26}{106}\right)+\left(\frac{33}{106}\right)-\left(\frac{21}{106}\right)=\left(\frac{38}{106}\right)$
5. $\frac{21}{33}$

121. a

3.10: Chapter 3 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.11: Chapter 3 References

3.2 Terminology

"Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

3.3 Independent and Mutually Exclusive Events

Lopez, Shane, Preety Sidhu. "U.S. Teachers Love Their Lives, but Struggle in the Workplace." Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/te...workplace.aspx (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

3.4 Two Basic Rules of Probability

DiCamillo, Mark, Mervin Field. "The File Poll." Field Research Corporation. Available online at www.field.com/fieldpollonline...rs/Rls2443.pdf (accessed May 2, 2013).

Rider, David, "Ford support plummeting, poll suggests," The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011..._suggests.html (accessed May 2, 2013).

"Mayor's Approval Down." News Release by Forum Research Inc. Available online at www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

"Roulette." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. "Language Use in the United States: 2007." United States Census Bureau. Available online at www.census.gov/hhes/socdemo/l...acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

3.5 Contingency Tables and Probability Trees

"Blood Types." American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-a...od/blood-types (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

"Human Blood Types." Unite Blood Services, 2011. Available online at https://www.vitalant.org/Donate/Blood-Donation/Donate-Blood-Overview.aspx (accessed May 2, 2013).

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).

Samuel, T. M. "Strange Facts about RH Negative Blood." eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_st...ive-blood.html (accessed May 2, 2013).

"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).

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at https://ropercenter.cornell.edu/?s=S...h+for+Datasets (accessed February 6, 2019).

3.11: Chapter 3 References is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.16: Reference** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.
- **Current page** has no license indicated.

CHAPTER OVERVIEW

4: The Normal Distribution

4.1: Introduction
4.2: The Standard Normal Distribution
4.3: Using the Normal Distribution
4.4: Chapter 4 Key Terms
4.5: Chapter 4 Review
4.6: Chapter 4 Formula Review
4.7: Chapter 4 Homework
4.8: Chapter 4 Solutions
4.9: Chapter 4 References

This page titled 4: The Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by .

4.1: Introduction



Figure 4.1.1 If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

The normal probability density function, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution.

The normal distribution is extremely important, but it cannot be applied to everything in the real world. Remember here that we are still talking about the distribution of population data. This is a discussion of probability and thus it is the population data that may be normally distributed, and if it is, then this is how we can find probabilities of specific events just as we did for population data that may be binomially distributed or Poisson distributed. This caution is here because in the next chapter we will see that the normal distribution describes something very different from raw data and forms the foundation of inferential statistics.

The normal distribution has two parameters (two numerical descriptive measures): the mean (μ) and the standard deviation (σ).

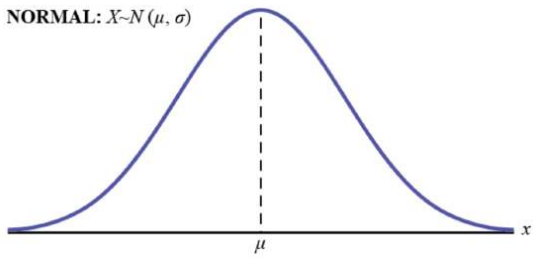


Figure 4.1.2

The curve is symmetric about a vertical line drawn through the mean, μ. The mean is the same as the median, which is the same as the mode, 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 normal curve; the curve becomes fatter and wider or skinnier and taller 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**.

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

4.2: The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called *z*-scores**. **A *z*-score is measured in units of the standard deviation.**

The mean for the standard normal distribution is zero, and the standard deviation is one. What this does is dramatically simplify the mathematical calculation of probabilities. The transformation $z = \frac{x-\mu}{\sigma}$ produces the distribution $Z \sim N(0,1)$. The value x in the given equation comes from a known normal distribution with known mean μ and known standard deviation σ. The z-score tells how many standard deviations a particular x is away from the mean.

Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z*-score for a particular x is:

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

The *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.

Suppose $X \sim N(5, 6)$. This says that X is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is two standard deviations (2σ) above or to the right of the mean $\mu = 5$.

Now suppose $x = 1$. Then: $z = \frac{x-\mu}{\sigma} = \frac{1-5}{6} = -0.67$ (rounded to two decimal places). This means that $\mathbf{x = 1}$ is 0.67 standard deviations $(\mathbf{-0.67\sigma})$ below or to the left of the mean $\mu = \mathbf{5}$.

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then **the Empirical Rule** states the following:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for $+1\sigma$ and -1σ are $+1$ and -1, respectively.
- The z-scores for $+2\sigma$ and -2σ are $+2$ and -2, respectively.
- The z-scores for $+3\sigma$ and -3σ are $+3$ and -3 respectively.

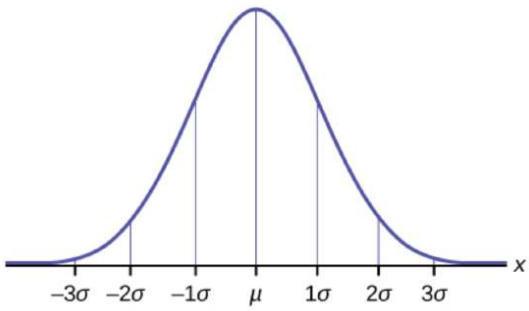


Figure 4.2.1

Example 4.2.1

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 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50-18 = 32$ and $50 + 18 = 68$ are within three standard deviations from the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

This page titled 4.2: The Standard Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by .

4.3: Using the Normal Distribution

The shaded area in the following graph indicates the area to the right of x. This area is represented by the probability $P(X > x)$. Normal tables - see Appendix A in Chapter 11.1 - provide the probability above a specific value such as x_1. This is the shaded part of the graph below.

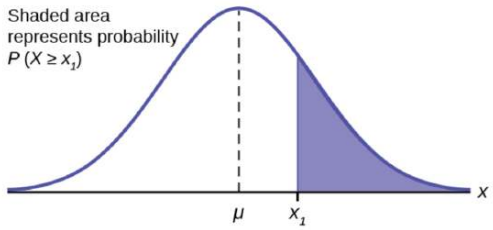


Figure 4.3.1

Because the normal distribution is symmetrical, if x_1 were the same distance to the left of the mean, the probability in the left tail (*below* that value), would be the same as the shaded area in the right tail as shown in the figure. In general, bear in mind that because of the symmetry of this distribution, one-half of the probability is to the right of the mean and one-half is to the left of the mean.

Calculations of Probabilities

Let's discuss how to find the probability of a specified region in a standard normal distribution. The shaded region in the figure below shows that the area between x_1 and x_2 is the probability as stated in the formula: $P(X_1 \leq X \leq X_2)$. In this case, suppose we have μ = 5 and σ = 2. Suppose that x_1 = 7 and x_2 = 9. We want to find the probability of a score falling between 7 and 9 on this distribution.

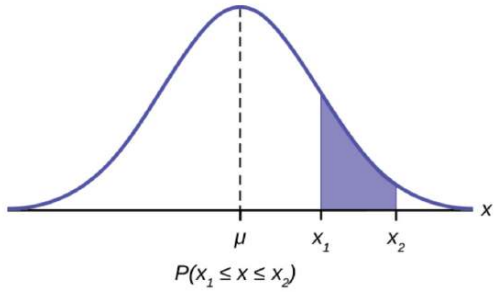


Figure 4.3.2

The solution is to convert the distribution we have with its mean and standard deviation to the Standard Normal Distribution. The Standard Normal has a random variable called Z. To compute probabilities, areas, for any normal distribution, we need only to convert the particular normal distribution to the standard normal distribution and look up the answer in the tables. As review, here again is the **standardizing formula**:

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

where z is the value on the standard normal distribution, x is the value from a normal distribution one wishes to convert to the standard normal, μ and σ are, respectively, the mean and standard deviation of that population. Note that the equation uses μ and σ which denotes population parameters. This is still dealing with probability so we always are dealing with the population, with **known** parameter values and a **known** distribution. It is also important to note that because the normal distribution is symmetrical it does not matter if the *z*-score is positive or negative when calculating a probability. One standard deviation to the left (negative *z*-score) covers the same area as one standard deviation to the right (positive *z*-score). This fact is why the Standard Normal tables do not provide areas for the left side of the distribution. Because of this symmetry, the *z*-score formula can also be written as:

$$Z = \frac{|x - \mu|}{\sigma}$$

where the vertical lines in the equation means the absolute value of the number.

Using this formula, we can determine that x_1 = 7 converts to a *z*-score of +1, and x_2 = 9 converts to a *z*-score of 2.

Then, using the z table, to find the probability of z = 1, go to the z column, reading down to 1.0 and then read at column 0. That number, 0.1587 is the probability of a score falling at or above z = 1. We repeat this process - using the z table - to find the probability of a score falling at or above z = 2, which is 0.0228. To obtain the shaded area that we wish to know about (as shown in Figure 4.3.2 above), we will need to subtract $0.1587 - 0.0228$ to obtain our final probability, $.1359$.

To compute probabilities, areas, for any normal distribution, we need only to convert the particular normal distribution to the standard normal distribution using the *z*-formula and look up the answer in the tables.

What the standardizing formula is really doing is computing the number of standard deviations x is from the mean of its own distribution. The standardizing formula and the concept of counting standard deviations from the mean is the secret of all that we will do in this statistics class. The reason this is true is that **all** of statistics boils down to variation, and the counting of standard deviations is a measure of variation.

This formula, in many disguises, will reappear over and over throughout this course.

Example 4.3.1

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.

Answer

a. Let x = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$.

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$

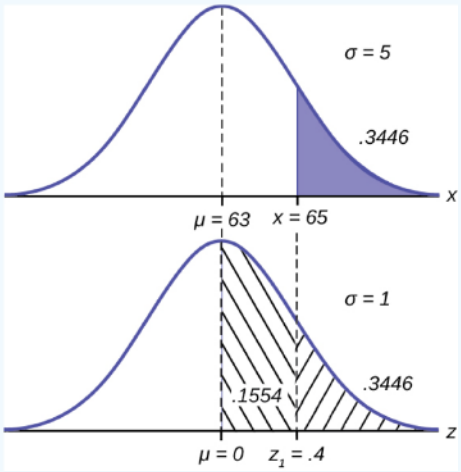


Figure 4.3.3

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{65 - 63}{5} = 0.4$$

Looking up this value of z in our z-table, $P(x \geq x_1) = P(z \geq z_1) = 0.3446$

https://biz.libretexts.org/@go/page/79019

The probability that any student selected at random scores more than 65 is 0.3446.

Answer

b.

$z = \frac{x-\mu}{\sigma} = \frac{85-63}{5} = 4.4$. The closest value in our *z*-table is 4.5, which has a probability of .00000340.

Therefore, the probability that one student scores *less* than 85 is approximately one or 100% (i.e., 1 - the probability of scoring *above* that value).

Exercise 4.3.1

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.

Example 4.3.2

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Answer

a. Let x = the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

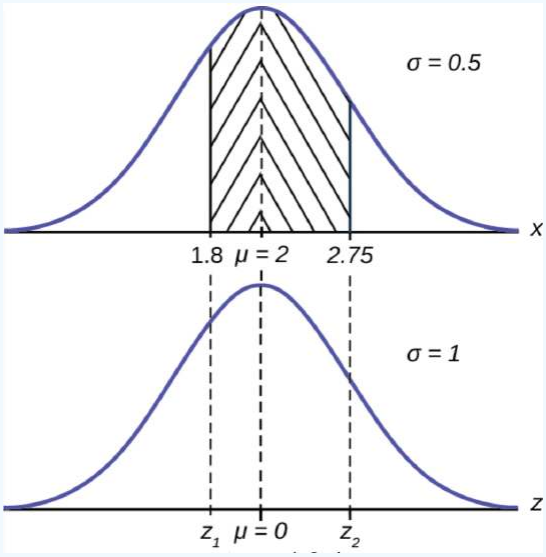


Figure 4.3.4

$P(1.8 \leq x \leq 2.75) = P(z_1 \leq z \leq z_2)$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Answer

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, first find the 25th percentile, k, where $P(x < k) = 0.25$.

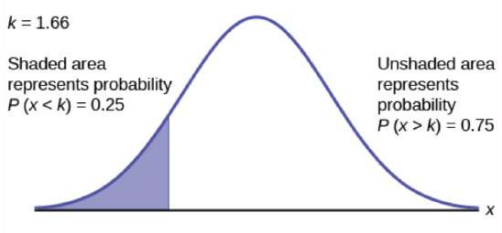


Figure 4.3.5

$f(Z) = 0.5 - 0.25 = 0.25$, therefore $z \approx -0.675$(or just 0.68 using the table) $z = \frac{x-\mu}{\sigma} = \frac{x-2}{0.5} = -0.675$, therefore $x = -0.675 * 0.5 + 2 = 1.66$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Exercise 4.3.2

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 golfer scored between 66 and 70.

Example 4.3.3

In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

Answer

a. 0.8186

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

Answer

b. 0.8413

Example 4.3.4

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.

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.

Answer

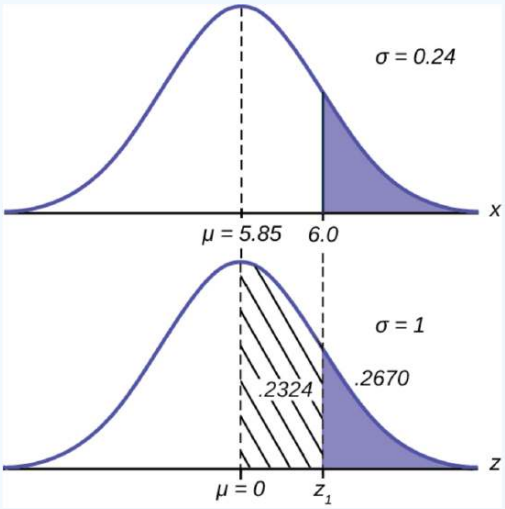


Figure 4.3.6

$$z_1 = \frac{6 - 5.85}{.24} = .625$$

$P(x \geq 6) = P(z \geq 0.625) = 0.2670$

b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.

Answer

$f(z) = \frac{0.20}{2} = 0.10,$ therefore $z \approx \pm 0.25$

$z = \frac{x-\mu}{\sigma} = \frac{x-5.85}{0.24} = \pm 0.25 \rightarrow \pm 0.25 \cdot 0.24 + 5.85 = (5.79, 5.91)$

This page titled 4.3: Using the Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by .

4.4: Chapter 4 Key Terms

Normal Distribution

a continuous random variable where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV, Z, is called the **standard normal distribution**.

Standard Normal Distribution

a continuous random variable $(RV) X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

***z*-score**

the linear transformation of the form $z = \frac{x-\mu}{\sigma}$ or written as $z = \frac{|x-\mu|}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the *z*-score of x. The *z*-score allows us to compare data that are normally distributed but scaled differently. A *z*-score is the number of standard deviations a particular x is away from its mean value.

This page titled 4.4: Chapter 4 Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by .

4.5: Chapter 4 Review

4.2 The Standard Normal Distribution

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0,1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of *z*-scores. Its mean is zero, and its standard deviation is one.

This page titled 4.5: Chapter 4 Review is shared under a CC BY license and was authored, remixed, and/or curated by .

4.6: Chapter 4 Formula Review

4.1 Introduction

$X \sim N(\mu, \sigma)$

$\mu =$ the mean; $\sigma =$ the standard deviation

4.2 The Standard Normal Distribution

$Z \sim N(0, 1)$

z = a standardized value (z-score)

mean = 0; standard deviation = 1

To find the kth percentile of X when the z-scores is known:
$k = \mu + (z)\sigma$

z-score: $z = \frac{x-\mu}{\sigma}$ or $z = \frac{|x-\mu|}{\sigma}$

$Z =$ the random variable for z-scores

$Z \sim N(0, 1)$

Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.

Standard Normal Distribution: $Z \sim N(0, 1)$.

4.6: Chapter 4 Formula Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

4.7: Chapter 4 Homework

4.2 The Standard Normal Distribution

1. A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. $X =$ _____.

2. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

3. $X \sim N(1, 2)$

$\sigma =$ _____

4. 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. $X =$ _____.

5. $X \sim N(-4, 1)$

What is the median?

6. $X \sim N(3, 5)$

$\sigma =$ _____

7. $X \sim N(-2, 1)$

$\mu =$ _____

8. What does a *z*-score measure?

9. What does standardizing a normal distribution do to the mean?

10. Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

11. What is the *z*-score of $x = 12$, if it is two standard deviations to the right of the mean?

12. What is the *z*-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

13. What is the *z*-score of $x = -2$, if it is 2.78 standard deviations to the right of the mean?

14. What is the *z*-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

15. Suppose $X \sim N(2, 6)$. What value of x has a *z*-score of 3?

16. Suppose $X \sim N(8, 1)$. What value of x has a *z*-score of –2.25?

17. Suppose $X \sim N(9, 5)$. What value of x has a *z*-score of –0.5?

18. Suppose $X \sim N(2, 3)$. What value of x has a *z*-score of –0.67?

19. Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

20. Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

21. Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

22. Suppose $X \sim N(-1, 2)$. What is the *z*-score of $x = 2$?

23. Suppose $X \sim N(12, 6)$. What is the *z*-score of $x = 2$?

24. Suppose $X \sim N(9, 3)$. What is the *z*-score of $x = 9$?

25. Suppose a normal distribution has a mean of 6 and a standard deviation of 1.5. What is the *z*-score of $x = 5.5$?

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

32. 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?

33. About what percent of x values lie between the second and third standard deviations (both sides)?

34. Suppose $X \sim N(15, 3)$. Between what x values does the middle 68.27% of the data lie?

35. Suppose $X \sim N(-3, 1)$. Between what x values does the middle 95.45% of the data lie?

36. Suppose $X \sim N(-3, 1)$. Between what x values does the middle 34.14% of the data lie?

37. About what percent of x values lie between the mean and three standard deviations?

38. About what percent of x values lie between the mean and one standard deviation?

39. About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

40. About what percent of x values lie between the first and third standard deviations (both sides)?

Use the following information to answer the next two exercises: The life of a particular gaming console is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. This gaming console is guaranteed for three years. We are interested in the length of time a given gaming console lasts.

41. Define the random variable X in words. $X =$ _____.

42. $X \sim$ _____(_____,_____)

4.3 Using the Normal Distribution

43. How would you represent the area to the left of 1 in a probability statement?

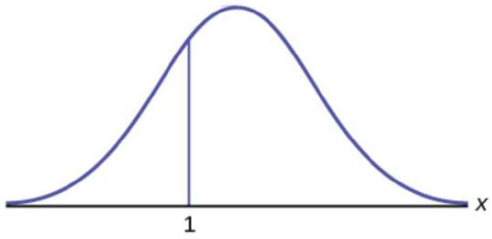


Figure 4.7.1

44. What is the area to the right of 1?

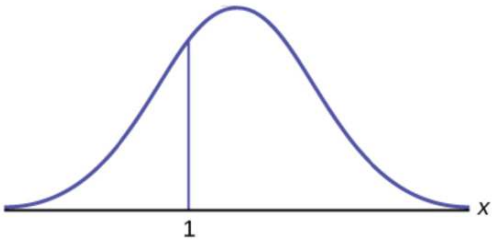


Figure 4.7.2

45. Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

46. How would you represent the area to the left of 3 in a probability statement?

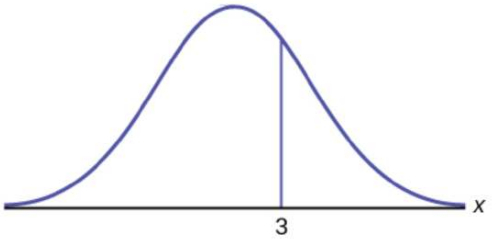


Figure 4.7.3

47. What is the area to the right of 3?

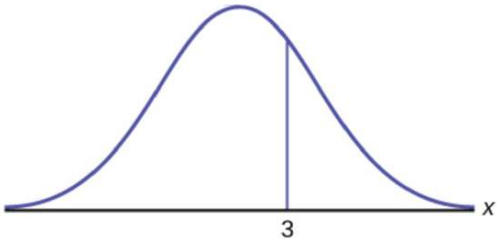


Figure 4.7.4

48. If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

49. If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Use the following information to answer the next two exercises: $X \sim N(54, 8)$

50. Find the probability that $x > 56$.

51. Find the probability that $x < 30$.

52. $X \sim N(6, 2)$

Find the probability that x is between 3 and 9.

53. $X \sim N(-3, 4)$

Find the probability that x is between 1 and 4.

54. $X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

55. *Use the following information to answer the next three exercises:* The life of a particular gaming console is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. This gaming console is guaranteed for three years. We are interested

in the length of time a given gaming console lasts. Find the probability that a gaming console will break down during the guarantee period.

1. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

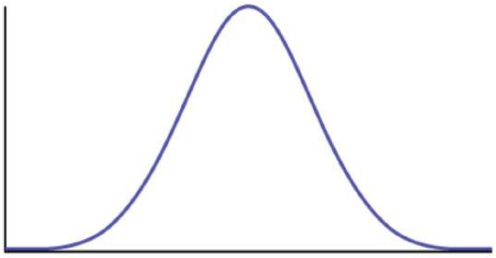


Figure 4.7.5

2. $P(0 < x <$ _____) = _____ (Use zero for the minimum value of x.)

56. Find the probability that a gaming console will last between 2.8 and six years.

1. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



Figure 4.7.6

2. P(_____ $< x <$ _____) = _____

Use the following information to answer the next two exercises: 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.

65. What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

66. What is the *z*-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

67. The length of time to find an open parking space at 9 A.M. 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 follows a normal distribution.
II. The data follows a right-skewed distribution.
III. The data follows a left-skewed distribution.

a. I only

b. II only
c. III only
d. I, II, and III

68. 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 a standard deviation, $\sigma = 3.89$ inches. For each of the following heights, calculate the z-score and interpret it using complete sentences.

1. 77 inches
2. 85 inches
3. If an NBA player reported his height had a *z*-score of 3.5, would you believe him? Explain your answer.

69. 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.

1. Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.
2. 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?

70. Kyle's doctor told him that the *z*-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$.

1. Which answer(s) is/are correct?
   - Kyle's systolic blood pressure is 175.
   - Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
   - Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
   - Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
2. Calculate Kyle's blood pressure.

71. 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 and interpret them.

1. 11 kg
2. 7.9 kg
3. 12.2 kg

72. 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 $\mu = 520$ and standard deviation $\sigma = 115$.

1. Calculate the *z*-score for an SAT score of 720. Interpret it using a complete sentence.
2. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
3. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. 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?

Use the following information to answer the next two exercises: 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.

73. What is the probability of spending more than two days in recovery?

a. 0.0582
b. 0.8447
c. 0.0553
d. 0.9418

Use the following information to answer the next three exercises: The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

74. Based upon the given information and numerically justified, would you be surprised if it took less than 1 minute to find a parking space?

a. Yes
b. No
c. Unable to determine

75. Find the probability that it takes at least 8 minutes to find a parking space.

a. 0.0001
b. 0.9270
c. 0.1862
d. 0.0668

76. Seventy percent of the time, it takes more than how many minutes to find a parking space?

a. 1.24
b. 2.41
c. 3.96
d. 6.05

77. According to a study done by Gettysburg students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let X = height of the individual.

1. $X \sim$ _____(_____,_____)
2. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
3. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
4. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

78. IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X = IQ of an individual.

1. $X \sim$ _____(_____,_____)
2. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
3. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.

79. The number of advertisements that a person in America views each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let X = number of advertisements seen per day.

1. $X \sim$ _____(_____,_____)
2. Find the probability that the number of advertisements seen per day is more than 40. Graph the situation. Shade in the area to be determined.
3. Find the maximum number for the lower quarter of number of advertisements seen per day. Sketch the graph and write the probability statement.

80. 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.

1. If X = distance in feet for a fly ball, then $X \sim$ _____(_____,_____)
2. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.

81. Four-year-olds average three hours a day of screen time (e.g., on electronic devices). Suppose that the standard deviation is 1.5 hours and the amount of screen time is normally distributed. We randomly select one four-year-old. We are interested in the amount of screen time the child experiences per day.

1. In words, define the random variable X.
2. $X \sim$ _____(_____,_____)
3. Find the probability that the child has less than one hour of screen time per day. Sketch the graph, and write the probability statement.
4. What percent of the children have over ten hours of screen time per day?
5. Seventy percent of the children have at least how much screen time per day?

82. In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let X = number of votes for President Clinton for an election district.

1. State the approximate distribution of X.
2. Is 1,956.8 a population mean or a sample mean? How do you know?
3. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
4. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
5. Find the third quartile for votes for President Clinton.

83. Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

1. In words, define the random variable X.
2. $X \sim$ _____(_____,_____)
3. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
4. Sixty percent of all trials of this type are completed within how many days?

84. Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

1. In words, define the random variable X.
2. $X \sim$ _____(_____,_____)
3. Find the percent of her laps that are completed in less than 130 seconds.
4. The fastest 3% of her laps are under _____.
5. The middle 80% of her laps are from _____ seconds to _____ seconds.

85. Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let X = time in line. Table 4.7.1 displays the ordered real data (in minutes):

1	5	5	6	7
4	5	6	7	11

Table 4.7.1

1. Calculate the sample mean and the sample standard deviation.
2. Construct a histogram.
3. Draw a smooth curve through the midpoints of the tops of the bars.
4. In words, describe the shape of your histogram and smooth curve.
5. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____)

6. Use the distribution in part e to calculate the probability that a person will wait fewer than 6 minutes.
7. Determine the cumulative relative frequency for waiting less than 6 minutes.
8. Why aren't the answers to part 6 and part 7 exactly the same?
9. Why are the answers to part 6 and part 7 as close as they are?
10. If only 5 customers has been surveyed rather than 10, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

86. Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

1. Ricardo's actual GPA is lower than Anita's actual GPA.
2. Ricardo is not passing because his *z*-score is zero.
3. Anita is in the 70th percentile of students at her college.

87. An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the *z*-scores first, and then use those to calculate the probability.

88. A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of $n = 100$ cars. Let X represent the number of defective cars in the sample. What can we say about values of x in regard to the Empirical Rule? Assume a normal distribution for the defective cars in the sample.

89. We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$. Solve the following:

1. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
2. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
3. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

90. A $1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

1. somewhere between 34 and 54 prizes.
2. somewhere between 54 and 64 prizes.
3. more than 64 prizes.

91. Facebook provides a variety of statistics on its Web site that detail the growth and popularity of the site.

On average, 28 percent of 18 to 34 year olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.

Use the normal distribution to determine the probability that half or more of a sample of *n* = 32 18-to-34 year olds checked Facebook before getting out of bed this the morning.

92. A hospital has 49 births in a year. It is considered equally likely that a birth be a boy as it is the birth be a girl.

1. What is the mean?
2. What is the standard deviation?
3. Use the normal distribution to find the probability that at least 23 of the 49 births were boys.

93. Historically, a final exam in a course is passed with a probability of 0.9. The exam is given to a group of 70 students.

1. What is the mean?
2. What is the standard deviation?
3. Use the normal distribution to find the probability that at least 60 of the students pass the exam?

94. A tree in an orchard has 200 oranges. Of the oranges, 40 are not ripe. Use the normal distribution to determine the probability a box containing 35 oranges has at most two oranges that are not ripe.

95. In a large city one in ten fire hydrants are in need of repair. If a crew examines 100 fire hydrants in a week, what is the probability they will find nine or fewer fire hydrants that need repair?

96. On an assembly line it is determined 85% of the assembled products have no defects. If one day 50 items are assembled, what is the probability at least 4 and no more than 8 are defective?

4.7: Chapter 4 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

4.8: Chapter 4 Solutions

1. ounces of water in a bottle

3. 2

5. –4

7. –2

9. The mean becomes zero.

11. $z = 2$

13. $z = 2.78$

15. $x = 20$

17. $x = 6.5$

19. $x = 1$

21. $x = 1.97$

23. $z = –1.67$

25. $z \approx –0.33$

27. 0.67, right

29. 3.14, left

31. about 68%

33. about 4%

35. between –5 and –1

37. about 50%

39. about 27%

41. The lifetime of a gaming console measured in years.

43. $P(x < 1)$

45. Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

47. $1–P(x < 3)$ or $P(x > 3)$

49. $1–0.543 = 0.457$

51. 0.0013

53. 0.1186

55.

1. Check student's solution.
2. 3, 0.1977

66. c

68.

1. Use the *z*-score formula. $z = –0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.
2. Use the *z*-score formula. $z = 1.5424$. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.

3. $\text{Height} = 79 + 3.5(3.89) = 92.615\text{inches}$, which is taller than 7 feet, 8 inches. There are very few NBA players this tall so the answer is no, not likely.

70.

1. iv
2. Kyle's blood pressure is equal to $125 + (1.75)(14) = 149.5$.

72. Let $X =$ an SAT math score and $Y =$ an ACT math score.

1. $X = 720, \frac{720-520}{15} = 1.74$ The exam score of 720 is 1.74 standard deviations above the mean of 520.
2. $z = 1.5$
   The math SAT score is $520 + 1.5(115) \approx 692.5$. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.
3. $\frac{X-\mu}{\sigma} = \frac{700-514}{117} \approx 1.59$, the z-score for the SAT. $\frac{Y-\mu}{\sigma} = \frac{30-21}{5.3} \approx 1.70$, the z-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher z-score).

75. d

77.

1. $X \sim N(66, 2.5)$
2. 0.5404
3. No, the probability that an Asian male is over 72 inches tall is 0.0082

79.

1. $X \sim N(36, 10)$
2. The probability that a person views more than 40 advertisements per day is 0.3446.
3. Approximately 25% of people view fewer than 29.26 advertisements per day.

81.

1. $X =$ number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
2. $X \sim N(3, 1.5)$
3. The probability that the child spends less than one hour a day unsupervised is 0.0918.
4. The probability that a child spends over ten hours a day unsupervised is less than 0.0001.
5. 2.22 hours

83.

1. $X =$ the distribution of the number of days a particular type of criminal trial will take
2. $X \sim N(21, 7)$
3. The probability that a randomly selected trial will last more than 24 days is 0.3336.
4. 22.77

85.

1. mean = 5.70, s = 2.54
2. Check student's solution.
3. Check student's solution.
4. Check student's solution.
5. $X \sim N(5.70, 2.54)$
6. 0.5478
7. The cumulative frequency for less than 6 minutes is 0.50.
8. The answers to part 6 and part 7 are not exactly the same, because the normal distribution is only an approximation to the real one.
9. The answers to part 6 and part 7 are close, because a normal distribution is an excellent approximation - but more so when the sample size is greater than 30.

10. The approximation would have been (even) less accurate, because the smaller sample size means that the data does not fit normal curve as well.

88.

- $n = 100; p = 0.1; q = 0.9$
- $\mu = np = (100)(0.10) = 10$
- $\sigma = \sqrt{npq} = \sqrt{(100)(0.1)(0.9)} = 3$

1. $z = \pm 1 : x_1 = \mu + z\sigma = 10 + 1(3) = 13$ and $x_2 = \mu - z\sigma = 10 - 1(3) = 7.68\%$ of the defective cars will fall between seven and 13.
2. $z = \pm 2 : x_1 = \mu + z\sigma = 10 + 2(3) = 16$ and $x_2 = \mu - z\sigma = 10 - 2(3) = 4.95\%$ of the defective cars will fall between four and 16
3. $z = \pm 3 : x_1 = \mu + z\sigma = 10 + 3(3) = 19$ and $x_2 = \mu - z\sigma = 10 - 3(3) = 1.99.7\%$ of the defective cars will fall between one and 19.

90.

- $n = 190; p = 1515 = 0.2; q = 0.8$
- $\mu = np = (190)(0.2) = 38$
- $\sigma = \sqrt{npq} = \sqrt{(190)(0.2)(0.8)} = 5.5136$

1. For this problem: $P(34 < x < 54) = 0.7641$
2. For this problem: $P(54 < x < 64) = 0.0018$
3. For this problem: $P(x > 64) = 0.0000012$ (approximately 0)

92.

1. 24.5
2. 3.5
3. Yes
4. 0.67

93.

1. 63
2. 2.5
3. Yes
4. 0.88

94. 0.02

95. 0.37

96. 0.50

4.8: Chapter 4 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

4.9: Chapter 4 References

4.2 The Standard Normal Distribution

"Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewre...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/digita...Group-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/d...s/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_o...ms_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

4.3 Using the Normal Distribution

"Naegele's rule." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).

"403: NUMMI." Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radi...sode/403/nummi (accessed May 14, 2013).

"Scratch-Off Lottery Ticket Playing Tips." WinAtTheLottery.com, 2013. Available online at www.winatthelottery.com/publi...partment40.cfm (accessed May 14, 2013).

"Smart Phone Users, By The Numbers." Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).

"Facebook Statistics." Statistics Brain. Available online at http://www.statisticbrain.com/facebo...tics/(accessed May 14, 2013).

4.9: Chapter 4 References is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.10: References** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.
- **Current page** has no license indicated.

CHAPTER OVERVIEW

5: The Central Limit Theorem

5.1: Introduction to the Central Limit Theorem

5.2: The Central Limit Theorem for Sample Means

5.3: Using the Central Limit Theorem

5.4: Chapter 5 Key Terms

5.5: Chapter 5 Review

5.6: Chapter 5 Formula Review

5.7: Chapter 5 Homework

5.8: Chapter 5 Solutions

5.9: Chapter 5 References

This page titled 5: The Central Limit Theorem is shared under a CC BY license and was authored, remixed, and/or curated by .

5.1: Introduction to the Central Limit Theorem

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **Central Limit Theorem**.

The **Central Limit Theorem** is one of the most powerful and useful ideas in all of statistics. The Central Limit Theorem is a theorem which means that it is NOT a theory or just somebody's idea of the way things work. As a theorem it ranks with the Pythagorean Theorem, or the theorem that tells us that the sum of the angles of a triangle must add to 180. These are facts of the ways of the world rigorously demonstrated with mathematical precision and logic. As we will see this powerful theorem will determine just what we can, and cannot say, in inferential statistics. The Central Limit Theorem is concerned with drawing finite samples of size n from a population with a known mean, μ, and a known standard deviation, σ. The conclusion is that if we collect samples of size n with a "large enough n," calculate each sample's mean, and create a histogram (distribution) of those means, then the resulting distribution will tend to have an approximate normal distribution.

The astounding result is that it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means tend to follow the normal distribution.



Figure 5.1.1 If you want to figure out the distribution of the change people carry in their pockets, using the Central Limit Theorem and assuming your sample is large enough, you will find that the distribution is normal. (credit: John Lodder)

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30, or ideally 100 or more). If the original population is far from normal, then more observations are needed for the sample means. Sampling is done randomly and with replacement in the theoretical model.

This page titled 5.1: Introduction to the Central Limit Theorem is shared under a CC BY license and was authored, remixed, and/or curated by .

5.2: The Central Limit Theorem for Sample Means

The sampling distribution is a theoretical distribution. It is created by taking many many samples of size n from a population. Each sample mean is then treated like a single observation of this new distribution, the sampling distribution. The genius of thinking this way is that it recognizes that when we sample we are creating an observation and that observation must come from some particular distribution. The Central Limit Theorem answers the question: from what distribution did a sample mean come? If this is discovered, then we can treat a sample mean just like any other observation and calculate probabilities about what values it might take on. We have effectively moved from the world of statistics where we know only what we have from the sample, to the world of probability where we know the distribution from which the sample mean came and the parameters of that distribution.

The reasons that one samples a population are obvious. The time and expense of checking every invoice to determine its validity or every shipment to see if it contains all the items may well exceed the cost of errors in billing or shipping. For some products, sampling would require destroying them, called destructive sampling. One such example is measuring the ability of a metal to withstand saltwater corrosion for parts on ocean-going vessels.

Sampling thus raises an important question, just which sample was drawn? Even if the sample were randomly drawn, there are theoretically an almost infinite number of samples. With just 100 items, there are more than 75 million unique samples of size five that can be drawn. If six are in the sample, the number of possible samples increases to just more than one billion. Of the 75 million possible samples, then, which one did you get? If there is variation in the items to be sampled, there will be variation in the samples. One could draw an "unlucky" sample and make very wrong conclusions concerning the population. This recognition that any sample we draw is really only one from a distribution of samples provides us with what is probably the single most important theorem is statistics, the **Central Limit Theorem**. Without the Central Limit Theorem it would be impossible to proceed to inferential statistics from simple probability theory. In its most basic form, the Central Limit Theorem states that **regardless** of the underlying probability density function of the population data, the theoretical distribution of the means of samples from the population will be normally distributed. In essence, this says that the mean of a sample should be treated like an observation drawn from a normal distribution. The Central Limit Theorem only holds if the sample size is "large enough" which has been shown to be only 30 observations or more.

Figure 5.2.1 graphically displays this very important proposition.

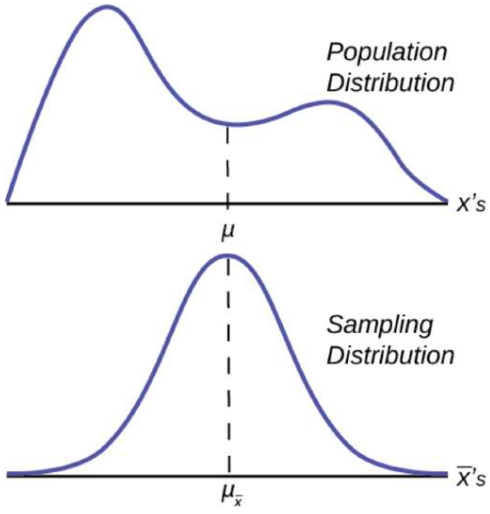


Figure 5.2.1

Notice that the horizontal axis in the top panel is labeled x. These are the individual observations of the population. This is the **unknown** distribution of the population values. The graph is purposefully drawn all squiggly to show that it does not matter just how odd ball it really is. Remember, we will never know what this distribution looks like, or its mean or standard deviation for that matter.

The horizontal axis in the bottom panel is labeled $\overline{x}$'s. This is the theoretical distribution called the sampling distribution of the means. Each observation on this distribution is a sample mean. All these sample means were calculated from individual samples with the same sample size. The theoretical sampling distribution contains all of the sample mean values from all the possible samples that could have been taken from the population. Of course, no one would ever actually take all of these samples, but if they did this is how they would look. And the Central Limit Theorem says that they will be normally distributed.

The Central Limit Theorem goes even further and tells us the mean and standard deviation of this theoretical distribution.

Table 5.2.1

Parameter	Population distribution	Sample	Sampling distribution of $\overline{x}$'s
Mean	μ	$\overline{x}$	$\mu_{\overline{x}}$ and $\mathrm{E}(\mu_{\overline{x}}) = \mu$
Standard deviation	σ	s	$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

The practical significance of The Central Limit Theorem is that now we can compute probabilities for drawing a sample mean, $\overline{X}$, in just the same way as we did for drawing specific observations, *X*s, when we knew the population mean and standard deviation and that the population data were normally distributed. The standardizing formula has to be amended to recognize that the mean and standard deviation of the sampling distribution, sometimes, called the standard error of the mean, are different from those of the population distribution, but otherwise nothing has changed. The new standardizing formula is

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Notice that $\mu_{\overline{x}}$ in the first formula has been changed to simply μ in the second version. The reason is that mathematically it can be shown that the expected value of $\mu_{\overline{x}}$ is equal to μ. This was stated in Table 5.2.1 above. Mathematically, the $E(x)$ symbol read the "expected value of x". This formula will be used in the next unit to provide estimates of the **unknown** population parameter μ.

This page titled 5.2: The Central Limit Theorem for Sample Means is shared under a CC BY license and was authored, remixed, and/or curated by .

5.3: Using the Central Limit Theorem

Examples of the Central Limit Theorem

Law of Large Numbers

The **law of large numbers** says that if you take samples of larger and larger size from any population, then the mean of the sampling distribution, $\mu_{\overline{x}}$ tends to get closer and closer to the true population mean, μ. From the Central Limit Theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation of the sampling distribution gets. (Remember that the standard deviation for the sampling distribution of $\overline{x}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\overline{x}$ must be closer to the population mean μ as n increases. 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.

This concept is so important and plays such a critical role in what follows it deserves to be developed further. Indeed, there are two critical issues that flow from the Central Limit Theorem and the application of the Law of Large numbers to it. These are:

1. the sampling distribution of means is normally distributed **regardless** of the underlying distribution of the population observations and
2. the standard deviation of the sampling distribution of means decreases as the size of the samples that were used to calculate the means for the sampling distribution increases.

It would seem counterintuitive that the population may have **any** distribution and the distribution of means coming from it would be normally distributed. With the use of computers, experiments can be simulated that show the process by which the sampling distribution changes as the sample size is increased. These simulations show visually the results of the mathematical proof of the Central Limit Theorem.

Here are three examples of very different population distributions and the evolution of the sampling distribution to a normal distribution as the sample size increases. The top panel in these cases represents the histogram for the original data. The three panels show the histograms for 1,000 randomly drawn samples for different sample sizes: $n = 10$, $n = 25$ and $n = 50$. As the sample size increases, and the number of samples taken remains constant, the distribution of the 1,000 sample means becomes closer to the smooth line that represents the normal distribution.

Figure 5.3.1 is for a normal distribution of individual observations and we would expect the sampling distribution to converge on the normal quickly. The results show this and show that even at a very small sample size the distribution is close to the normal distribution.

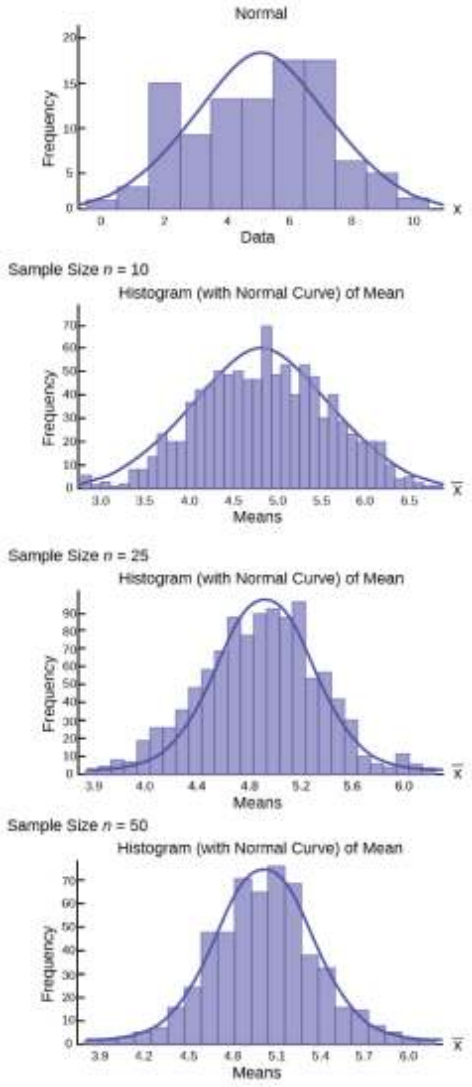


Figure 5.3.1

Figure 5.3.2 is a skewed distribution. This last one could be an exponential, geometric, or binomial with a small probability of success creating the skew in the distribution. For skewed distributions our intuition would say that this will take larger sample sizes to move to a normal distribution and indeed that is what we observe from the simulation. Nevertheless, at a sample size of 50, not considered a very large sample, the distribution of sample means has very decidedly gained the shape of the normal distribution.

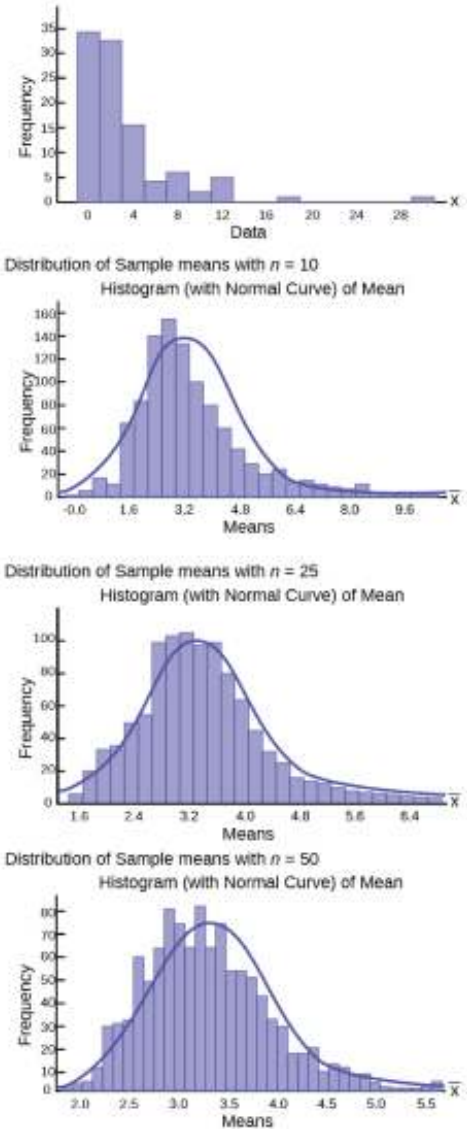


Figure 5.3.2

Figure 5.3.3 is a uniform distribution which, a bit amazingly, quickly approached the normal distribution even with only a sample of 10.

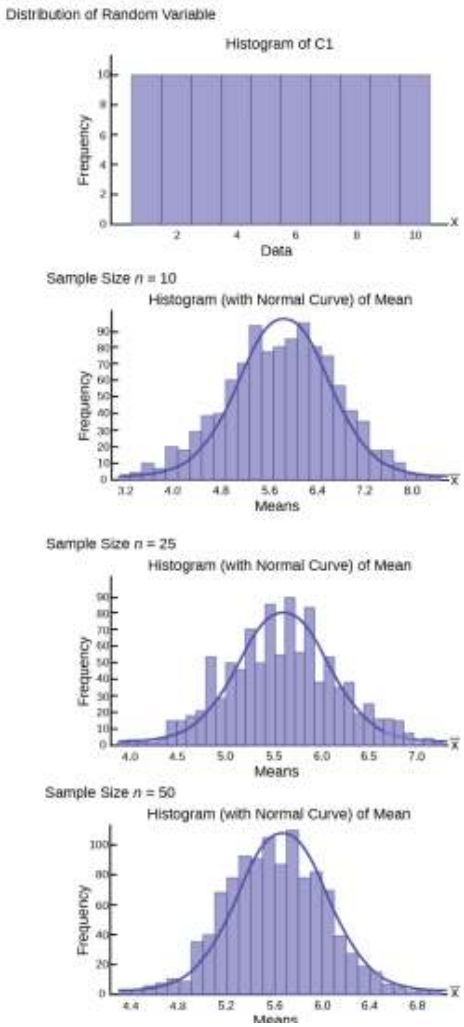


Figure 5.3.3

The Central Limit Theorem provides more than the proof that the sampling distribution of means is normally distributed. It also provides us with the mean and standard deviation of this distribution. Further, as discussed above, the expected value of the mean, $\mu_{\overline{x}}$, is equal to the mean of the population of the original data which is what we are interested in estimating from the sample we took. We have already inserted this conclusion of the Central Limit Theorem into the formula we use for standardizing from the sampling distribution to the standard normal distribution. And finally, the Central Limit Theorem has also provided the standard deviation of the sampling distribution, $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, and this is critical to have to calculate probabilities of values of the new random variable, $\overline{x}$.

Figure 5.3.4 shows three sampling distributions. The mean has been marked on the horizontal axis of the $\overline{x}$'s. Remember that the standard deviation of the sampling distribution is the original standard deviation of the population, divided by the sample size. We have already seen that as the sample size increases the sampling distribution becomes closer and closer to the normal distribution. As this happens, the standard deviation of the sampling distribution changes in another way; the standard deviation decreases as n increases. At very very large n, the standard deviation of the sampling distribution becomes very small and at infinity it collapses on top of the population mean. This is what it means that the expected value of $\mu_{\overline{x}}$ is the population mean, μ.

At non-extreme values of n, this relationship between the standard deviation of the sampling distribution and the sample size plays a very important part in our ability to estimate the parameters we are interested in.

The only change that was made in Figure 5.3.4 is the sample size that was used to get the sample means for each distribution. As the sample size increases, n goes from 10 to 30 to 50, the standard deviations of the respective sampling distributions decrease because the sample size is in the denominator of the standard deviations of the sampling distributions.

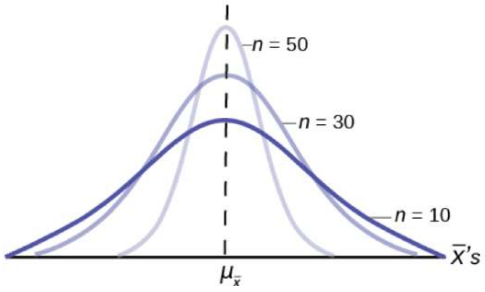


Figure 5.3.4

The implications for this are very important. Figure 5.3.5 shows the effect of the sample size on the confidence we will have in our estimates. These are two sampling distributions from the same population. One sampling distribution was created with samples of size 10 and the other with samples of size 50. All other things constant, the sampling distribution with sample size 50 has a smaller standard deviation that causes the graph to be higher and narrower. The important effect of this is that for the same probability of one standard deviation from the mean, this distribution covers much less of a range of possible values than the other distribution. One standard deviation is marked on the $\overline{x}$ axis for each distribution. This is shown by the two arrows that are plus or minus one standard deviation for each distribution. For the sampling distribution with the smaller sample size, the possible range of $\overline{x}$ values is much greater. A simple question is, would you rather have a sample mean from the narrow, tight distribution, or the flat, wide distribution as the estimate of the population mean? Your answer tells us why people intuitively will always choose data from a large sample rather than a small sample. The sample mean they are getting is coming from a more compact distribution. This concept will be the foundation for what will be called level of confidence in the next unit.

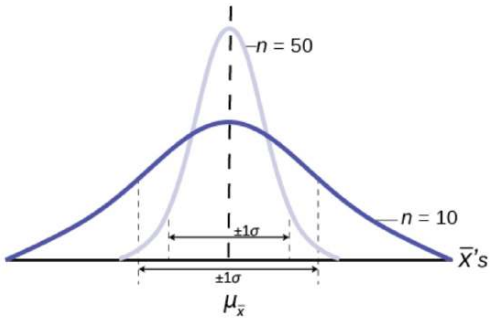


Figure 5.3.5

This page titled 5.3: Using the Central Limit Theorem is shared under a CC BY license and was authored, remixed, and/or curated by .

5.4: Chapter 5 Key Terms

Central Limit Theorem

Given a random variable with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, $\overline{X}$. If the size (n) of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Mean

A number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Normal Distribution

Notation: $X \sim N(\mu, \sigma)$ for a continuous random variable, where μ is the mean of the distribution and σ is the standard deviation. If $\mu = 0$ and $\sigma = 1$, the random variable, z, is called the **standard normal distribution**.

Sampling Distribution

Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

Standard Error of the Mean

The standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$

This page titled 5.4: Chapter 5 Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by .

5.5: Chapter 5 Review

5.2 The Central Limit Theorem for Sample Means

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

5.3 Using the Central Limit Theorem

The Central Limit Theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample size you take from a population, the closer the sample mean $\overline{x}$ gets to μ.

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

5.6: Chapter 5 Formula Review

5.1 The Central Limit Theorem for Sample Means

The Central Limit Theorem for Sample Means:

$\overline{x} \sim N\left(\mu_{\overline{x}}, \frac{\sigma}{\sqrt{n}}\right)$

Mean $\overline{x}$: $\mu_{\overline{x}}$

Standard Error of the Mean (Standard Deviation): $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem for Sample Means z-score: $z = \frac{\overline{x}-\mu_{\overline{x}}}{\sigma_x} = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$

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

5.7: Chapter 5 Homework

5.3 Using the Central Limit Theorem

Use the following information to answer the next eight exercises: The length of time a particular smartphone's battery lasts follows an normal distribution with a mean of 10 months and a standard deviation of 10 months. A sample of 64 of these smartphones is taken.

1. a. What is the mean of the sampling distribution?

b. What is the standard deviation of the sampling distribution?

2. What is the distribution for the length of time one battery lasts?

3. What is the distribution for the mean length of time 64 batteries last?

4. What is the distribution for the total length of time 64 batteries last?

5. Find the probability that the sample mean is between seven and 11.

6. Find the 80th percentile for the total length of time 64 batteries last.

7. Find the IQR for the mean amount of time 64 batteries last.

8. Find the middle 80% for the total amount of time 64 batteries last.

9. A population has a mean of 25 and a standard deviation of 2. If it is sampled repeatedly with samples of size 49, what is the mean and standard deviation of the sample means?

10. A population has a mean of 48 and a standard deviation of 5. If it is sampled repeatedly with samples of size 36, what is the mean and standard deviation of the sample means?

11. A population has a mean of 90 and a standard deviation of 6. If it is sampled repeatedly with samples of size 64, what is the mean and standard deviation of the sample means?

12. A population has a mean of 120 and a standard deviation of 2.4. If it is sampled repeatedly with samples of size 40, what is the mean and standard deviation of the sample means?

13. A population has a mean of 17 and a standard deviation of 1.2. If it is sampled repeatedly with samples of size 50, what is the mean and standard deviation of the sample means?

14. A population has a mean of 17 and a standard deviation of 0.2. If it is sampled repeatedly with samples of size 16, what is the expected value of the mean and the standard deviation of the sample means?

15. A population has a mean of 38 and a standard deviation of 3. If it is sampled repeatedly with samples of size 48, what is the expected value of the mean and the standard deviation of the sample means?

16. A population has a mean of 14 and a standard deviation of 5. If it is sampled repeatedly with samples of size 60, what is the expected value of the mean and the standard deviation of the sample means?

17. A fishing boat has 1,000 fish on board, with an average weight of 120 pounds and a standard deviation of 6.0 pounds. If sample sizes of 50 fish are checked, what is the probability the fish in a sample will have mean weight within 2.0 pounds of the true mean of the population?

18. An experimental garden has 500 sunflower plants. The plants are being treated so they grow to unusual heights. The average height is 9.3 feet with a standard deviation of 0.5 foot. If sample sizes of 60 plants are taken, what is the probability the plants in a given sample will have an average height within 0.1 foot of the true mean of the population?

19. A company has 800 employees. The average number of workdays between absence for illness is 123 with a standard deviation of 14 days. Samples of 50 employees are examined. What is the probability a sample has a mean of workdays with no absence for illness of at least 124 days?

20. Cars pass an automatic speed check device that monitors 2,000 cars on a given day. This population of cars has an average speed of 67 miles per hour with a standard deviation of 2 miles per hour. If samples of 30 cars are taken, what is the probability a given sample will have an average speed within 0.50 mile per hour of the population mean?

21. A town keeps weather records. From these records it has been determined that it rains on an average of 37% of the days each year. If 30 days are selected at random from one year, what is the probability that at least 5 and at most 11 days had rain?

22. A maker of yardsticks has an ink problem that causes the markings to smear on 4% of the yardsticks. The daily production run is 2,000 yardsticks. What is the probability if a sample of 100 yardsticks is checked, there will be ink smeared on at most 4 yardsticks?

23. A school has 300 students. Usually, there are an average of 21 students who are absent. If a sample of 30 students is taken on a certain day, what is the probability that at most 2 students in the sample will be absent?

24. A college gives a placement test to 5,000 incoming students each year. On the average 1,213 place in one or more developmental courses. If a sample of 50 is taken from the 5,000, what is the probability at most 12 of those sampled will have to take at least one developmental course?

5.2 The Central Limit Theorem for Sample Means

25. Previously, Gettysburg College statistics students estimated that the amount of change statistics students carry is normally distributed with a mean of $0.88 and a standard deviation of $0.31. Suppose that we randomly pick 25 statistics students.

1. In words, X = _____
2. $X \sim$ _____(_____,_____)
3. In words, $\overline{x}$ = _____
4. $\overline{x} \sim$ _____ (_____, _____)
5. Find the probability that an individual had between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
6. 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.
7. Explain why there is a difference in part e and part f.

26. 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.

1. If $\overline{x}$ = average distance in feet for 49 fly balls, then $\overline{x} \sim$ _____(_____,_____)
2. 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.
3. Find the 80th percentile of the distribution of the average of 49 fly balls.

27. 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.

1. In words, X = _____
2. In words, $\overline{x}$ = _____
3. $\overline{x} \sim$ _____(_____,_____)
4. 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.
5. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

28. Suppose that a category of world-class runners are known to run a marathon (26 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.

1. $\overline{x} \sim$ _____(_____,_____)

2. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
3. Find the 80th percentile for the average of these 49 marathons.
4. Find the median of the average running times.

29. The length of songs in a collector's Spotify collection is normally distributed with a mean of 2.75 minutes and a standard deviation of 0.43 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

1. In words, X = _____
2. $X \sim$ _____
3. In words, $\overline{x}$ = _____
4. $\overline{x} \sim$ _____(_____,_____)
5. Find the first quartile for the average song length.
6. The IQR (interquartile range) for the average song length is from _____–_____.

30. In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

1. In words, X = _____
2. In words, $\overline{x}$ = _____
3. $\overline{x} \sim$ _____(_____,_____)
4. The IQR for $\overline{x}$ is from _____ acres to _____ acres.

31. Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

1. When the sample size is large, the mean of $\overline{x}$ is approximately equal to the population mean of X.
2. When the sample size is large, the sampling distribution of $\overline{x}$ is approximately normally distributed.
3. When the sample size is large, the standard deviation of $\overline{x}$ is approximately the same as the standard deviation of X.

32. The percent of calories from protein that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten. Suppose that 16 individuals are randomly chosen. Let $\overline{x}$ = average percent of calories from protein.

a. $\overline{x} \sim$ _____(_____, _____)
b. For the group of 16, find the probability that the average percent of calories from protein 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 calories from protein.

33. 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.

1. In words, X = _____
2. In words, $\overline{x}$ = _____
3. $\overline{x} \sim$ _____(_____,_____)
4. How is it possible for the standard deviation to be greater than the average?
5. 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?

34. Which of the following is NOT TRUE about the distribution for averages?

1. The mean, median, and mode are equal.
2. The area under the curve is one.
3. The curve never touches the x-axis.
4. The curve is skewed to the right.

35. The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. 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. $\overline{x} \sim N(4.59, 0.10)$

b. $\overline{x} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$

c. $\overline{x} \sim N\left(4.59, \frac{16}{0.10}\right)$

d. $\overline{x} \sim N\left(4.59, \frac{\sqrt{16}}{0.10}\right)$

5.3 Using the Central Limit Theorem

36. A large population of 5,000 students take a practice test to prepare for a standardized test. The population mean is 140 questions correct, and the standard deviation is 80. What size samples should a researcher take to get a distribution of means of the samples with a standard deviation of 10?

37. A large population has skewed data with a mean of 70 and a standard deviation of 6. Samples of size 100 are taken, and the distribution of the means of these samples is analyzed.

1. Will the distribution of the means be closer to a normal distribution than the distribution of the population?
2. Will the mean of the means of the samples remain close to 70?
3. Will the distribution of the means have a smaller standard deviation?
4. What is that standard deviation?

38. A researcher is looking at data from a large population with a standard deviation that is much too large. In order to concentrate the information, the researcher decides to repeatedly sample the data and use the distribution of the means of the samples? The first effort used sample sized of 100. But the standard deviation was about double the value the researcher wanted. What is the smallest size samples the researcher can use to remedy the problem?

39. A researcher looks at a large set of data, and concludes the population has a standard deviation of 40. Using sample sizes of 64, the researcher is able to focus the mean of the means of the sample to a narrower distribution where the standard deviation is 5. Then, the researcher realizes there was an error in the original calculations, and the initial standard deviation is really 20. Since the standard deviation of the means of the samples was obtained using the original standard deviation, this value is also impacted by the discovery of the error. What is the correct value of the standard deviation of the means of the samples?

40. A population has a standard deviation of 50. It is sampled with samples of size 100. What is the variance of the means of the samples?

41. A company has 1,000 employees. The average number of workdays between absence for illness is 80 with a standard deviation of 11 days. Samples of 80 employees are examined. What is the probability a sample has a mean of workdays with no absence for illness of at least 78 days and at most 84 days?

42. Trucks pass an automatic scale that monitors 2,000 trucks. This population of trucks has an average weight of 20 tons with a standard deviation of 2 tons. If a sample of 50 trucks is taken, what is the probability the sample will have an average weight within one-half ton of the population mean?

43. A town keeps weather records. From these records it has been determined that it rains on an average of 12% of the days each year. If 30 days are selected at random from one year, what is the probability that at most 3 days had rain?

44. A maker of greeting cards has an ink problem that causes the ink to smear on 7% of the cards. The daily production run is 500 cards. What is the probability that if a sample of 35 cards is checked, there will be ink smeared on at most 5 cards?

45. A school has 500 students. Usually, there are an average of 20 students who are absent. If a sample of 30 students is taken on a certain day, what is the probability that at least 2 students in the sample will be absent?

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

5.8: Chapter 5 Solutions

1. a. 10

b. 1.25

3. $N(10, 1.25)$

5. 0.7799

7. 1.675

9. Mean = 25, standard deviation = 0.29

10. Mean = 48, standard deviation = 0.83

11. Mean = 90, standard deviation = 0.75

12. Mean = 120, standard deviation = 0.38

13. Mean = 17, standard deviation = 0.17

14. Expected value = 17, standard deviation = 0.05

15. Expected value = 38, standard deviation = 0.43

16. Expected value = 14, standard deviation = 0.65

17. .9818

18. .8788

19. .3050

20. .8294

21. .4884

22. .5000

23. .4761

24. .5160

25.

1. X = amount of change students carry
2. $X \sim N(0.88, 0.31)$
3. $\overline{x}$ = average amount of change carried by a sample of 25 students.
4. $\overline{x} \sim N(0.88, 0.062)$
5. $.2543$
6. $.8753$
7. The distributions are different. Part 1 is normal for individual observations and part 2 is normal for sample means.

27.

1. length of time for an individual to complete IRS form 1040, in hours.
2. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.
3. $N(10.53, \frac{1}{3})$
4. Yes. I would be surprised, because the probability is almost 0.
5. No. I would not be totally surprised because the probability is .2296.

29.

1. the length of a song, in minutes, in the collection
2. $N(2, 0.43)$

3. the average length, in minutes, of the songs from a sample of five albums from the collection
4. $N(2.75, 0.066)$
5. 2.706 minutes
6. 0.088 minutes

31.

1. True. The mean of a sampling distribution of the means is approximately the mean of the population.
2. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
3. False. The standard deviation of the sampling distribution of the means will decrease as the sample size increases.

33.

1. X = the yearly income of someone in a third world country
2. the average salary from samples of 1,000 residents of a third world country
3. $\overline{x} \sim N\left(2000, \frac{8000}{\sqrt{1000}}\right)$
4. Very wide differences in data values can have averages smaller than standard deviations.
5. The distribution of the sample mean will have higher probabilities closer to the population mean.
   $P(2000 < \overline{x} < 2100) = .1537$
   $P(2100 < \overline{x} < 2200) = .1317$

35. b

36. 64

37.

1. Yes
2. Yes
3. Yes
4. 0.6

38. 400

39. 2.5

40. 25

41. .947

42. .9232

43. .648

44. .101

45. .273

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

5.9: Chapter 5 References

5.2 The Central Limit Theorem for Sample Means

Baran, Daya. “20 Percent of Americans Have Never Used Email.”WebGuild, 2010. Available online at http://www.webguild.org/20080519/20-...ver-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.

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

CHAPTER OVERVIEW

6: Confidence Intervals

6.1: Introduction

6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size

6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

6.4: A Confidence Interval for A Population Proportion

6.5: Chapter 6 Key Terms

6.6: Chapter 6 Review

6.7: Chapter 6 Formula Review

6.8: Chapter 6 Homework

6.9: Chapter 6 Solutions

6.10: Chapter 6 References

This page titled 6: Confidence Intervals is shared under a CC BY license and was authored, remixed, and/or curated by .

6.1: Introduction



Figure 6.1.1 Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Suppose you were 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 attempted. In this case, you would have obtained a point estimate for the true proportion the parameter *P*.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called **confidence intervals**. What statistics provides us beyond a simple average, or point estimate, is an estimate to which we can attach a probability of accuracy, what we will call a **confidence level**. We make inferences with a known level of probability.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's *t* distribution, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that 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, $\bar{x}$, and the sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

$\bar{x}$ and s are each called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include the unknown population parameter.

Suppose, for the iTunes example, we know the sample mean $\bar{x} = 2$, the sample standard deviation is $s = 1$, and our sample size is 100. Then, by the central limit theorem, the standard deviation of the sampling distribution of the sample means is

$\frac{s}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$.

The **empirical rule**, which applies to the normal distribution, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is $(2)(0.1) = 0.2$. Therefore, the sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

The population mean μ is estimated 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, we have 95% confidence that μ is between $\overline{x} - 0.2$ and $\overline{x} + 0.2$.

In this case, with 95% probability 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 say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is [1.8, 2.2]. Please note that we talked in terms of 95% confidence using the empirical rule. The empirical rule for two standard deviations is only *approximately* 95% of the probability under the normal distribution. To be precise, two standard deviations under a normal distribution is actually 95.44% of the probability. To calculate the exact 95% confidence level we would use 1.96 standard deviations.

The 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 (100% minus 95% = 5%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ.

For the confidence interval for a mean the formula would be:

$$\mu = \overline{x} \pm z_{\frac{\alpha}{2}} * s/\sqrt{n}$$

Or written another way as:

$$\overline{x} - z_{\frac{\alpha}{2}} * s/_{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}} * s/\sqrt{n}$$

Where $\overline{x}$ is the sample mean. $z_{\frac{\alpha}{2}}$ is determined by the level of confidence (1-α) desired by the analyst, and $s/\sqrt{n}$ is the standard deviation of the sampling distribution for means given to us by the Central Limit Theorem.

This page titled 6.1: Introduction is shared under a CC BY license and was authored, remixed, and/or curated by .

6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size

A confidence interval for a population mean with a known population standard deviation is based on the conclusion of the Central Limit Theorem that the sampling distribution of the sample means follow an approximately normal distribution.

Calculating the Confidence Interval

Consider the standardizing formula for the sampling distribution developed in the discussion of the Central Limit Theorem:

$$z_1 = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Notice that μ is substituted for $\mu_{\overline{x}}$ because we know that the expected value of $\mu_{\overline{x}}$ is μ from the Central Limit Theorem and $\sigma_{\overline{x}}$ is replaced with $\sigma/\sqrt{n}$, also from the Central Limit Theorem.

In this formula we know $\overline{x}$, $\sigma_{\overline{x}}$ and n, the sample size. (In actuality we do not know the population standard deviation, but we do have a point estimate for it, s, from the sample we took. More on this later.) What we do not know is μ or z_1. We can solve for either one of these in terms of the other. Solving for μ in terms of z_1 gives:

$$\mu = \overline{x} \pm z_1\sigma/\sqrt{n}$$

Remembering that the Central Limit Theorem tells us that the distribution of the $\overline{x}$'s, the sampling distribution for means, is normal, and that the normal distribution is symmetrical, we can rearrange terms thus:

$$\overline{x} - z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$$

This is the formula for a confidence interval for the mean of a population.

Notice that $z_{\frac{\alpha}{2}}$ has been substituted for z_1 in this equation. This is where a choice must be made by the statistician. The analyst must decide the level of confidence they wish to impose on the confidence interval. α is the probability that the interval will not contain the true population mean. The confidence level is defined as $(1-\alpha)$. $z_{\frac{\alpha}{2}}$ is the number of standard deviations $\overline{x}$ lies from the mean with a certain probability. If we chose $z = 1.96$ we are asking for the 95% confidence interval because we are setting the probability that the true mean lies within the range at 0.95. If we set z at 1.645 we are asking for the 90% confidence interval because we have set the probability at 0.90. These numbers can be verified by consulting the z-table (see Appendix A). Divide either (1-0.95) or (1-0.90) in half and find that probability inside the body of the table. Then read on the corresponding top and left margins the number of standard deviations it takes to get this level of probability.

In reality, we can set whatever level of confidence we desire simply by changing the $z_{\frac{\alpha}{2}}$ value in the formula. It is the analyst's choice. Convention in business research and most social sciences sets confidence intervals at either 90, 95, or 99 percent levels. Levels less than 90% are considered of little value. The level of confidence of a particular interval estimate is equal to $(1-\alpha)$.

A good way to see the development of a confidence interval is to graphically depict the solution to a problem requesting a confidence interval. This is presented in the figure below for the example in the introduction concerning the number of downloads from iTunes. That case was for a 95% confidence interval, but other levels of confidence could have just as easily been chosen depending on the need of the analyst. However, the level of confidence MUST be pre-set and not subject to revision as a result of the calculations.

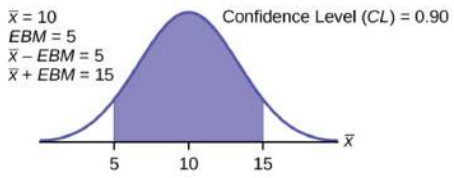


Figure 6.2.1

For this example, let's say we know that the actual population mean number of iTunes downloads is 2.1. The true population mean falls within the range of the 95% confidence interval. There is absolutely nothing to guarantee that this will happen. **Further, if the**

true mean falls outside of the interval we will never know it. We must always remember that we will never ever know the true mean. Statistics simply allows us, with a given level of probability (confidence), to say that the true mean is within the range calculated.

Changing the Confidence Level or Sample Size

Here again is the formula for a confidence interval for an unknown population mean assuming we know the population standard deviation:

$$\overline{x} - z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$$

It is clear that the confidence interval is driven by two things, the chosen level of confidence, $z_{\frac{\alpha}{2}}$, and the standard deviation of the sampling distribution. The standard deviation of the sampling distribution is further affected by two things, the standard deviation of the population and the sample size we chose for our data. Here we wish to examine the effects of each of the choices we have made on the calculated confidence interval, the confidence level and the sample size.

For a moment we should ask just what we desire in a confidence interval. Our goal was to estimate the population mean from a sample. We have forsaken the hope that we will ever find the true population mean, and population standard deviation for that matter, for any case except where we have an extremely small population and the cost of gathering the data of interest is very small. In all other cases we must rely on samples. With the Central Limit Theorem we have the tools to provide a meaningful confidence interval with a given level of confidence, meaning a known probability of being wrong. By meaningful confidence interval we mean one that is useful. Imagine that you are asked for a confidence interval for the ages of your classmates. You have taken a sample and find a mean of 19.8 years. You wish to be very confident so you report an interval between 9.8 years and 29.8 years. This interval would certainly contain the true population mean and have a very high confidence level. However, it hardly qualifies as meaningful. The very best confidence interval is narrow while having high confidence. There is a natural tension between these two goals. The higher the level of confidence, the wider the confidence interval as the case of the students' ages above. We can see this tension in the equation for the confidence interval.

$$\mu = \overline{x} \pm z_{\alpha}\left(\frac{\sigma}{\sqrt{n}}\right)$$

The confidence interval widens as confidence level increases, because z_{α} will become larger. Sample size also plays an important role in the width of the confidence interval. Notice that the sample size, n, shows up in the denominator of the standard deviation of the sampling distribution. Therefore, as the sample size increases, the standard deviation of the sampling distribution decreases, making the confidence interval narrower. Again, we see the importance of having large samples for our analysis here, although we then face a second constraint, the cost of gathering more data.

Calculating the Confidence Interval: An Alternative Approach

Another way to approach confidence intervals is through the use of something called the error bound. The **error bound** - also known as the **margin of error** - gets its name from the recognition that it provides the boundary of the interval derived from the standard error of the sampling distribution. In the equations above it is seen that the interval is simply the estimated mean, sample mean, plus or minus something. That something is the error bound and is driven by the probability we desire to maintain in our estimate, z_{α}, times the standard deviation of the sampling distribution. The error bound for a mean is given the name, **error bound mean**, or EBM.

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. Here, the margin of error (EBM) is the error bound for a population mean. The sample mean $\overline{x}$ is the **point estimate** of the unknown population mean μ.

The confidence interval estimate will have the form:

[point estimate - error bound, point estimate + error bound] or, in symbols, $[\overline{x} - EBM, \overline{x} + EBM]$

The mathematical formula for this confidence interval is:

$\overline{x} - z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$

The margin of error depends in part on the **confidence level** (abbreviated 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 the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is the probability that the interval does not contain the unknown population parameter. Mathematically, $1 - \alpha = CL$.

A confidence interval for a population mean with a *known* standard deviation is based on the fact that the sampling distribution of 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 $EBM = 5$.

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 $\alpha = .10$ in both tails, or 5% in each tail, of the normal distribution.

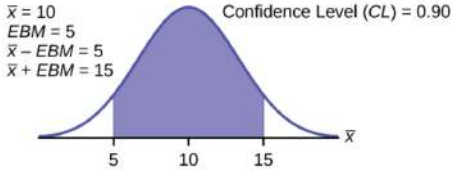


Figure 6.2.2

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 puts 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 the sampling distribution for means which we studied with the Central Limit Theorem and is, $\frac{\sigma}{\sqrt{n}}$.

Calculating the Confidence Interval Using EBM

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:

- Calculate the sample mean $\overline{x}$ from the sample data. Remember, in this section we know the population standard deviation σ.
- Find the *z*-score from the standard normal table that corresponds to the confidence level desired.
- Calculate the error bound.
- Construct the confidence interval.
- Write a sentence that interprets the confidence interval in the context of the situation in the problem.

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the *z*-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM 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 $Z \sim N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1–\alpha$, so 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 $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$. The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1–0.025 = 0.975$.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a standard normal probability table. We will see later that we can use a different probability table, the Student's t-distribution, for finding the number of standard deviations of commonly used levels of confidence.

Calculating the Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation \sigma is known is

- $EBM = (z_{\frac{\alpha}{2}}) \left(\frac{\sigma}{\sqrt{n}}\right)$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(\overline{x} - EBM, \overline{x} + EBM)$ or the formula: $\overline{x} - z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$

The graph gives a picture of the entire situation.

$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1$.

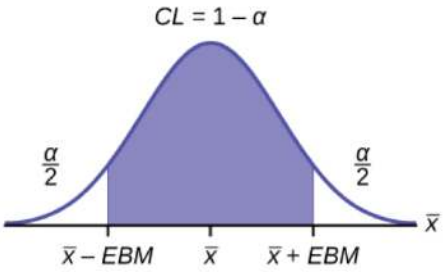


Figure 6.2.3

Example 6.2.1

Suppose we are interested in the mean scores on an exam. A random sample of 136 scores is taken and gives a sample mean (sample mean score) of 68. In this example we have the unusual knowledge that the population standard deviation is 3 points. (Do not count on knowing the population parameters outside of textbook examples!) Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

To find the confidence interval, you need the sample mean, $\overline{x}$, and the EBM.

$\overline{x} = 68$

$EBM = \left(z_{\frac{\alpha}{2}}\right) \left(\frac{\sigma}{\sqrt{n}}\right)$

$\sigma = 3$; $n = 136$; the confidence level is 90% $(CL = 0.90)$

$CL = 0.90$ so $\alpha = 1–CL = 1–0.90 = 0.10$

$\frac{\alpha}{2} = 0.05, z_{\frac{\alpha}{2}} = z_{0.05}$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1–0.05 = 0.95$.

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

Because the common levels of confidence in the social sciences are 90%, 95% and 99%, it will not be long until you become familiar with the numbers 1.645, 1.96, and 2.576.

$EBM = (1.645) \left(\frac{3}{\sqrt{136}}\right) = 0.423$

$\overline{x} - EBM = 68 - 0.423 = 67.577$

$\overline{x} + EBM = 68 - 0.423 = 68.423$

The 90% confidence interval is [67.577, 68.423].

Interpretation: We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.58 and 68.42.

Example 6.2.2

Suppose we change the original problem in Example 6.2.1 by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

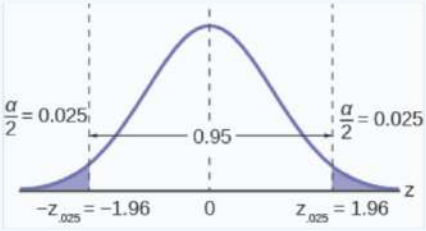


Figure 6.2.4

$$\mu = \overline{x} \pm z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$\mu = 68 \pm 1.96 \left(\frac{3}{\sqrt{136}} \right)$$

$$67.50 \leq \mu \leq 68.50$$

$\sigma = 3$; $n = 36$; the confidence level is 95% ($CL = 0.95$).

$CL = 0.95$ so $\alpha = 1 – CL = 1 – 0.95 = 0.05$

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Notice that the EBM is larger and the confidence interval is wider for a 95% confidence level, than it was in the original problem.

Comparing the results

Compared to the 90% confidence interval, the 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, 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, the confidence interval necessarily needs to be wider. This demonstrates a very important principle of confidence intervals. There is a trade off between the level of confidence and the width of the interval. Our desire is to have a narrow confidence interval, huge wide intervals provide little information that is useful. But we would also like to have a high level of confidence in our interval. This demonstrates that we cannot have both.

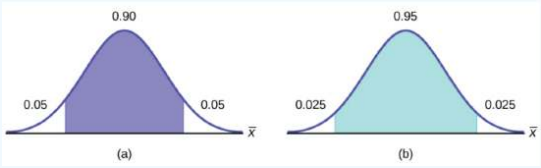


Figure 6.2.5

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level makes the confidence interval wider.
- Decreasing the confidence level makes the confidence interval narrower.

And again here is the formula for a confidence interval for an unknown mean assuming we have the population standard deviation:

$$\overline{x} - z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$$

The standard deviation of the sampling distribution was provided by the Central Limit Theorem as $\sigma/\sqrt{n}$. While we infrequently get to choose the sample size it plays an important role in the confidence interval. Because the sample size is in the denominator of the equation, as n increases it causes the standard deviation of the sampling distribution to decrease and thus the width of the confidence interval to decrease. We have met this before as we reviewed the effects of sample size on the Central Limit Theorem. There we saw that as n increases the sampling distribution narrows.

Example 6.2.3

Suppose we change the original problem in Example 6.2.1 once again to see what happens to the confidence interval if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the confidence interval if we increase the sample size and use $n = 200$ instead of $n = 136$? What happens if we decrease the sample size to $n = 115$ instead of $n = 136$?

Answer

Solution A

$\mu = \overline{x} \pm z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}}\right)$

$\mu = 68 \pm 1.645 \left(\frac{3}{\sqrt{200}}\right)$

$67.65 \leq \mu \leq 68.35$

If we **increase** the sample size n to 200, we **decrease** the width of the confidence interval relative to the original sample size of 136 observations.

Answer

Solution B

$\mu = \overline{x} \pm z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}}\right)$

$\mu = 68 \pm 1.645 \left(\frac{3}{\sqrt{115}}\right)$

$67.54 \leq \mu \leq 68.46$

If we **decrease** the sample size n to 115, we **increase** the width of the confidence interval by comparison to the original sample size of 136 observations.

Summary: Effect of Changing the Sample Size

- Increasing the sample size makes the confidence interval narrower.
- Decreasing the sample size makes the confidence interval wider.

We have already seen this effect when we reviewed the effects of changing the size of the sample, n, on the Central Limit Theorem. Before we saw that as the sample size increased the standard deviation of the sampling distribution decreases. This was why we choose the sample mean from a large sample as compared to a small sample, all other things held constant.

Thus far we assumed that we knew the population standard deviation. This will virtually never be the case. We will have the sample standard deviation, s, however. This is a point estimate for the population standard deviation and can be substituted into the formula for confidence intervals for a mean under certain circumstances. We just saw the effect the sample size has on the width of confidence interval and the impact on the sampling distribution for our discussion of the Central Limit Theorem. We can invoke

this to substitute the point estimate for the standard deviation if the sample size is large "enough". Simulation studies indicate that 30 observations or more will be sufficient to eliminate any meaningful bias in the estimated confidence interval.

Example 6.2.4

Spring break can be a very expensive holiday. A sample of 120 students is surveyed, and the average amount spent by students on travel and beverages is $593.84. The sample standard deviation is approximately $369.34.

Construct a 95% confidence interval for the population mean amount of money spent by spring breakers.

Answer

We begin with the confidence interval for a mean. We use the formula for a mean because the random variable is dollars spent and this is a continuous random variable. The point estimate for the population standard deviation, *s*, has been substituted for the true population standard deviation.

$$\mu = \overline{x} \pm \left[z_{(\alpha/2)} \frac{s}{\sqrt{n}} \right]$$

Substituting the values into the formula, we have:

$$\mu = 593.84 \pm \left[1.96 \frac{369.34}{\sqrt{120}} \right]$$

$z_{(a/2)}$ is found on the standard normal table by looking up 0.025 in the body of the table and finding the number of standard deviations on the side and top of the table; 1.96. The solution for the interval is thus:

$$\mu = 593.84 \pm 66.09 = [527.75, 659.93]$$

$$\$527.75 \leq \mu \leq \$659.93$$

This page titled 6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size is shared under a CC BY license and was authored, remixed, and/or curated by .

6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

In practice, we rarely know the population **standard deviation**. In earlier problems we have encountered, when the sample size was large, this did not present a problem. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. For example, this is what we did in Example 6.2.4. The point estimate for the standard deviation, s, was substituted in the formula for the confidence interval for the population standard deviation. In this case, the 120 observations were well above the suggested 100 observations to eliminate any bias from a small sample. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (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 produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's *t*-distribution**. The name comes from the fact that Gosset wrote under the pen name "A Student."

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 observations, but the two distributions only become interchangeable once n = 100 or more.

If you draw a simple random sample of size n from a population 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 in standard deviation units $\overline{x}$ is from its mean μ. For each sample size n, there is a different Student's *t*-distribution.

The **degrees of freedom (*df*)**, $n - 1$, come from the calculation of the sample standard deviation s. Remember when we first calculated a sample standard deviation we divided the sum of the squared deviations by $n - 1$, but we used n deviations 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 (*df*) in recognition that one is lost in the calculations. The effect of losing a degree of freedom is that the *t*-value increases and the confidence interval increases in width.

Properties of the Student's *t*-Distribution

- The graph for the Student's *t*-distribution is similar to the standard normal curve and at infinite degrees of freedom it is the normal distribution. You can confirm this by reading the bottom line at infinite degrees of freedom for a familiar level of confidence, e.g. at column 0.05, 95% level of confidence, we find the *t*-value of 1.96 at infinite degrees of freedom.
- The mean for the Student's *t*-distribution is zero and the distribution is symmetric about zero, again like the standard normal distribution.
- 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. This assumption comes from the central limit theorem because the individual observations in this case are the $\overline{x}$s of the sampling distribution. The size of the underlying population is generally not relevant unless it is very small. If it is normal then the assumption is met and doesn't need discussion.

A probability table for the Student's *t*-distribution (see Appendix A) is used to identify *t*-values at various commonly-used levels of confidence. The table gives *t*-scores that correspond to the confidence level (column) and degrees of freedom (row). Notice that at the bottom the table will show the *t*-value for infinite degrees of freedom. Mathematically, as the degrees of freedom increase, the t distribution approaches the standard normal distribution. You can find familiar *z*-values by looking in the relevant alpha column and reading value in the last row.

A Student's *t* table (see Appendix A) gives *t*-scores given the degrees of freedom and either a specified level of confidence or the right-tailed probability (depending on which row across the top you are using).

The Student's *t*-distribution has one of the most desirable properties of the normal: it is symmetrical. What the Student's *t*-distribution does is spread out the horizontal axis so it takes a larger number of standard deviations to capture the same amount of probability. In reality there are an infinite number of Student's *t*-distributions, one for each *df*. As the sample size increases, the Student's *t*-distribution become more and more like the normal distribution. When the sample size reaches 100 the normal distribution is usually substituted for the Student's *t* because they are so much alike. This relationship between the Student's *t*-distribution and the normal distribution is shown in Figure 6.3.1.

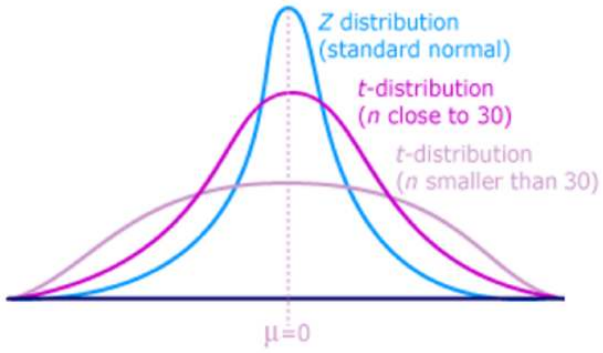


Figure 6.3.1

Restating the formula for a confidence interval for the mean for cases when the sample size is smaller than 100 and we do not know the population standard deviation, σ:

$$\overline{x} - t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right)$$

Here the point estimate of the population standard deviation, s has been substituted for the population standard deviation, σ, and $t_{\frac{\alpha}{2},df}$ has been substituted for $z_{\frac{\alpha}{2}}$. The notation for *df* is placed in the general formula in recognition that there are many Student *t*-distributions, one for each *df*. For this type of problem, the degrees of freedom is $df = n - 1$, where n is the sample size. To look up a probability in the Student's *t*-table we have to know the degrees of freedom in the problem.

Example 6.3.1

The average earnings per share (EPS) for 10 industrial stocks randomly selected from those listed on the Dow-Jones Industrial Average was found to be $\overline{x} = 1.85$ with a standard deviation of $s = 0.395$. Calculate a 99% confidence interval for the average EPS of all the industrials listed on the Dow Jones.

$$\overline{x} - t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right)$$

Answer

In this case, we will use the Student's *t*-distribution because we do not know the population standard deviation and the sample is small, less than 100.

To find the appropriate *t*-value requires two pieces of information: the level of confidence desired and the degrees of freedom. The question asked for a 99% confidence level. The tails, thus, need to have .005 probability each, $\alpha/2$. The degrees of freedom for this type of problem is $n - 1 = 9$.

From the Student's *t*-table, at the row marked 9 and column marked for $\alpha = .005$, is the number of standard deviations to capture 99% of the probability in the *t*-distribution, and it is 3.250. Remembering that the Student's t is symmetrical, this *t*-value is both plus or minus - appearing on each side of the mean in the distribution.

Inserting these values into the formula gives the result.

$$\mu = \overline{x} \pm t_{.005,9}\frac{s}{\sqrt{n}} = 1.85 \pm 3.250\frac{0.395}{\sqrt{10}} = 1.85 \pm 0.406$$

$$1.44 \leq \mu \leq 2.26$$

We can interpret this CI by saying: With 99% confidence, the average EPS of all the industries listed on DJIA is between \$1.44 and \$2.26.

Exercise 6.3.1

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 with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

This page titled 6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case is shared under a CC BY license and was authored, remixed, and/or curated by .

6.4: A Confidence Interval for A Population Proportion

During an election year, we see articles in the newspaper 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.

Investors in the stock market are interested in the true proportion of stocks that go up and down 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 go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval for a population proportion is similar to that for the population mean, but the formulas are a bit different although conceptually identical. While the formulas are different, they are based upon the same mathematical foundation given to us by the Central Limit Theorem. Because of this we will see the same basic format using the same three pieces of information: the sample value of the parameter in question, the standard deviation of the relevant sampling distribution, and the number of standard deviations we need to have the confidence in our estimate that we desire.

How do you know you are dealing with a proportion problem? First, the underlying distribution has a **binary random variable**. (There is no mention of a mean or average.) To form a sample proportion, take x, the number of successes (or other observations of interest) and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is the sample proportion,

$P' = \frac{x}{n}$

x = the **number** of successes in the sample

n = the size of the sample

P' = the **estimated proportion** of successes or sample proportion of successes (P' is a **point estimate** for P, the true population proportion, and thus $1 - P$ is the probability of a "failure" in any one trial.)

The population standard deviation of this estimate is equal to:

$$\sigma_{P'} = \sqrt{\frac{P(1-P)}{n}}$$

The confidence interval for a population proportion, therefore, becomes:

$$P' \pm \left[z_{(\frac{a}{2})} \sqrt{\frac{P'(1-P')}{n}} \right]$$

$z_{(\frac{a}{2})}$ is set according to our desired degree of confidence and $\sqrt{\frac{P'(1-P')}{n}}$ is the estimated standard deviation of the sampling distribution (using the sample proportion as the point estimate for the population one).

Example 6.4.1

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 sampled, 421 responded yes - 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.

Answer

Let x = the number of people in the sample who have cell phones. x is binomial: the random variable is binary, people either have a cell phone or they do not.

To calculate the confidence interval, we must find P'.

$n = 500$

$x =$ the number of people who own cell phones in the sample $= 421$

$P' = \frac{x}{n} = \frac{421}{500} = 0.842$

Since the requested confidence level is $CL = 0.95$, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$, and $(\frac{\alpha}{2}) = 0.025$.

Therefore, $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$. This can be found using the *z* table in Appendix A. This can also be found in the Student's *t*-table at the 0.025 column and the infinity degrees of freedom row, because at infinite degrees of freedom the Student's *t*-distribution becomes identical to the standard normal distribution, Z.

The confidence interval for the true population proportion is

$$P' - z_\alpha \sqrt{\frac{P'(1-P')}{n}} \leq P \leq P' + z_\alpha \sqrt{\frac{P'(1-P')}{n}}$$

Substituting in the values from above we find the confidence interval is: $0.810 \leq P \leq 0.874$

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.

Exercise 6.4.1

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Example 6.4.2

The Dundee Dog Training School has a larger than average proportion of clients who compete in competitive professional events. A confidence interval for the population proportion of dogs that compete in professional events from 150 different training schools is constructed. The lower limit is determined to be 0.08 and the upper limit is determined to be 0.16. Determine the level of confidence used to construct the interval of the population proportion of dogs that compete in professional events.

Answer

We begin with the formula for a confidence interval for a proportion because the random variable is binary; either the client competes in professional competitive dog events or they don't.

$$P' \pm \left[z_{(\frac{\alpha}{2})} \sqrt{\frac{P'(1-P')}{n}} \right]$$

Next we find the sample proportion:

$$P' = \frac{0.08 + 0.16}{2} = 0.12$$

The $\pm$ that makes up the confidence interval is thus 0.04.

$0.12 + 0.04 = 0.16$ and $0.12 - 0.04 = 0.08$, which are the boundaries of the confidence interval.

Finally, we solve for z.

$\left[z \cdot \sqrt{\frac{0.12(1-0.12)}{150}} \right] = 0.04, \ \mathbf{therefore\ z = 1.51}$

And then look up the probability for 1.51 standard deviations on the standard normal table.

$P(z = 1.51) = 0.4345, P(z) \cdot 2 = 0.8690$ or 86.90%.

Example 6.4.3

A financial officer for a company wants to estimate the percent of accounts receivable that are more than 30 days overdue. He surveys 500 accounts and finds that 300 are more than 30 days overdue. Compute a 90% confidence interval for the true percent of accounts receivable that are more than 30 days overdue, and interpret the confidence interval.

Answer

$x = 300$ and $n = 500$, so $P' = \frac{x}{n} = \frac{300}{500} = 0.60$

Since confidence level = 0.90, then $\alpha = 1 - \text{ confidence level } = (1 - 0.90) = 0.10$, and $(\frac{\alpha}{2}) = 0.05$

Therefore, $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$. This *z*-value can be found using a standard normal probability table. The student's *t*-table can also be used by checking the table at the 0.05 tail probability column and reading at the line for infinite degrees of freedom. (The *t*-distribution is the normal distribution at infinite degrees of freedom. This is a handy trick to remember in finding *z*-values for commonly used levels of confidence.)

We use this formula for a confidence interval for a proportion:

$$P' - z_{\alpha/2}\sqrt{\frac{P'(1-P')}{n}} \leq P \leq P' + z_{\alpha/2}\sqrt{\frac{P'(1-P')}{n}}$$

Substituting in the values from above we find the confidence interval for the true population proportion is $0.564 \leq P \leq 0.636$

Interpretation

We estimate with 90% confidence that the true percent of all accounts receivable overdue by 30+ days is between 56.4% and 63.6%. Alternate wording: We estimate with 90% confidence that between 56.4% and 63.6% of all accounts are 30+ days overdue.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of accounts receivable that are overdue 30 days.

Exercise 6.4.2

A student polls his school to see if students in the school district are for or against new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

This page titled 6.4: A Confidence Interval for A Population Proportion is shared under a CC BY license and was authored, remixed, and/or curated by .

6.5: Chapter 6 Key Terms

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Confidence Level (CL)

the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Degrees of Freedom (*df*)

the number of objects in a sample that are free to vary

Error Bound for a Population Mean (EBM)

the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Proportion (EBP)

the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Normal Distribution

notation: $X\sim N(\mu,\sigma)$. If $\mu=0$ and $\sigma=1$, the RV is called **the standard normal distribution**.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean, on average; notation: s for sample standard deviation and σ for population standard deviation

Student's *t*-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of this random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero.
- It approaches the standard normal distribution as n get larger.
- There is a "family" of *t* distributions: each representative of the family is completely defined by the number of degrees of freedom, which depends upon the application for which the t is being used.

6.5: Chapter 6 Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

- **8.6: Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

- **Current page** is licensed CC BY 4.0.

6.6: Chapter 6 Review

6.3 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, *s*, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when *s* is used as an estimate. In this case, the Student's t-distribution is much better. Define a *t*-score using the following formula:

$t = \frac{\overline{x}-\mu}{s/\sqrt{n}}$

The *t*-score follows the Student's *t*-distribution with $n-1$ degrees of freedom.

The confidence interval under this distribution is calculated with

$$\overline{x} - t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right)$$

where $t_{\frac{\alpha}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use the *t* table in Appendix A to find $t_{\frac{\alpha}{2}}$ for a given α and *df*.

6.4 A Confidence Interval for A Population Proportion

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let P' represent the sample proportion, x/n, where x represents the number of successes and n represents the sample size. Then the confidence interval for a population proportion is given by the following formula:

$$P' - z_{\alpha/2}\sqrt{\frac{P'(1-P')}{n}} \leq P \leq P' + z_{\alpha/2}\sqrt{\frac{P'(1-P')}{n}}$$

6.6: Chapter 6 Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.7: Chapter 6 Formula Review

6.3 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

s = the standard deviation of sample values.

$t = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ is the formula for the *t*-score which measures how far away a sample mean is from the population mean in the Student's *t*-distribution

$df = n - 1$; the degrees of freedom for a Student's *t*-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a Student's t-distribution with df degrees of freedom

The general form for a confidence interval for a single mean, population standard deviation unknown, and sample size less than 100 is given by:

$$\overline{x} - t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + t_{\frac{\alpha}{2},df}\left(\frac{s}{\sqrt{n}}\right)$$

6.4 A Confidence Interval for A Population Proportion

$P' = \frac{x}{n}$ where x represents the number of successes in a sample and n represents the sample size. The variable P' is the sample proportion and serves as the point estimate for the true population proportion, P.

The confidence interval for the true population proportion is given by the formula:

$$P' - z_\alpha\sqrt{\frac{P'(1-P')}{n}} \leq P \leq P' + z_\alpha\sqrt{\frac{P'(1-P')}{n}}$$

6.7: Chapter 6 Formula Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.8: Chapter 6 Homework

6.3 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

Use the following information to answer the next five exercises. 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.

1. Identify the following:

1. $\overline{x}$ =_____
2. s_x =_____
3. n =_____
4. $n - 1$ =_____

2. Define the random variables X and $\overline{x}$ in words.

3. Which distribution should you use for this problem?

4. Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

5. Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: 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 32.6 hours. Assume that the underlying population distribution is normal.

6. Identify the following:

1. $\overline{x}$ =_____
2. s_x =_____
3. n =_____
4. $n - 1$ =_____

7. Define the random variable X in words.

8. Define the random variable $\overline{x}$ in words.

9. Which distribution should you use for this problem?

10. Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

11. Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in the table below are the result of a random survey of 39 national flags 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	Freq.
1	1
2	7
3	18
4	7
5	6

Table 6.8.1

12. Calculate the following:

1. $\overline{x}$ =_____
2. s_x =_____
3. n =_____

13. Define the random variable $\overline{x}$ in words.

14. What is $\overline{x}$ estimating?

15. Is σ_x known?

16. As a result of your answer to the prior question, state the exact distribution you will use for calculating the confidence interval.

Construct a 95% confidence interval for the true mean number of colors on national flags.

17. How much area is in both tails (combined)?

18. How much area is in each tail?

19. Calculate the following:

1. lower limit
2. upper limit
3. error bound

20. The 95% confidence interval is_____.

21. Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

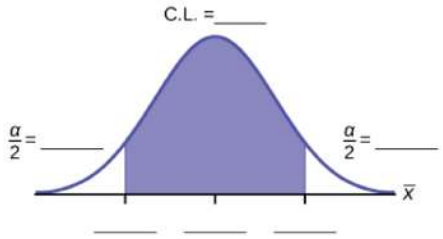


Figure 6.8.1

22. In one complete sentence, explain what the interval means.

23. Using the same $\overline{x}$, s_x, and level of confidence, suppose that *n* were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

24. Using the same $\overline{x}$, s_x, and $n = 39$, how would the error bound change if the confidence level were reduced to 90%? Why?

6.4 A Confidence Interval for A Population Proportion

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

25. 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?

26. 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?

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They 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.

27. Identify the following:

1. x = _____
2. n = _____
3. P' = _____

28. Define the random variables X and P' in words.

29. Which distribution should you use for this problem?

30. 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 error bound.

31. List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: 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.

32. We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and P' in words.

33. Which distribution should you use for this problem?

34. Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

35. Suppose we want to lower the sampling error. What is one way to accomplish that?

36. The sampling error given in the survey is ±2%. Explain what the ±2% means.

Use the following information to answer the next five exercises: 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.

37. Define the random variable X in words.

38. Define the random variable P' in words.

39. Which distribution should you use for this problem?

40. Construct a 90% confidence interval, and state the confidence interval and the error bound.

41. What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 36 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

42. What is being counted?

43. In words, define the random variable X.

44. Calculate the following:

1. x = _____
2. n = _____
3. P' = _____

45. State the estimated distribution of X. $X \sim$_____

46. Define a new random variable P'. What is P' estimating?

47. In words, define the random variable P'.

48. Construct a 95% confidence interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

49. How much area is in both tails (combined)?

50. How much area is in each tail?

51. Calculate the following:

1. lower limit
2. upper limit
3. error bound

52. The 95% confidence interval is _____.

53. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

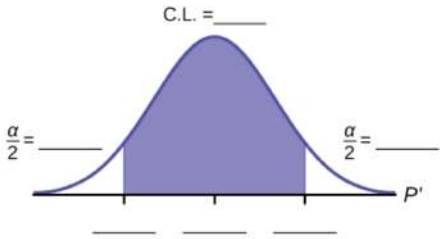


Figure 6.8.2

54. In one complete sentence, explain what the interval means.

55. Using the same P' and level of confidence, suppose that n were increased to 200. Would the error bound become larger or smaller? How do you know?

56. Using the same P' and $n = 100$, how would the error bound change if the confidence level were increased to 99%? Why?

57. If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

Use the following information to answer the next five exercises: 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.

58. Identify the following:

1. $\overline{x}$ = _____
2. σ = _____
3. n = _____

59. In words, define the random variables X and $\overline{x}$.

60. Which distribution should you use for this problem?

61. Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

62. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Use the following information to answer the next seven exercises: 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.

63. Identify the following:

1. $\overline{x}$ = _____

2. σ = _____
3. n = _____

64. In words, define the random variables X and $\overline{x}$.

65. Which distribution should you use for this problem?

66. Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

67. If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

68. If the Census did another survey, kept the error bound the same, and surveyed only 150 people instead of 200, what would happen to the level of confidence? Why?

69. Suppose the Census needed to be 99% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: 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.

70. Identify the following:

1. $\overline{x}$ = _____
2. σ = _____
3. n = _____

71. In words, define the random variable X.

72. In words, define the random variable $\overline{x}$.

73. Which distribution should you use for this problem?

74. 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 error bound.

75. 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 error bound.

76. In complete sentences, explain why the confidence interval in Question #75 is larger than in Question #74.

77. In complete sentences, give an interpretation of what the interval in Question #75 means.

78. What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

79. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: 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 twenty-five 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. Let X = the age of a Winter Foothill College student.

80. $\overline{x}$ = _____

81. n = _____

82. _____ = 15

83. In words, define the random variable $\overline{x}$.

84. What is $\overline{x}$ estimating?

85. Is σ_x known?

86. As a result of your answer to Question #83, state the exact distribution you will use for calculating the confidence interval.

Construct a 95% confidence interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

87. How much area is in both tails (combined)? α =_____

88. How much area is in each tail? $\frac{\alpha}{2}$ =_____

89. Identify the following specifications:

1. lower limit
2. upper limit
3. error bound

90. The 95% confidence interval is:_____.

91. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

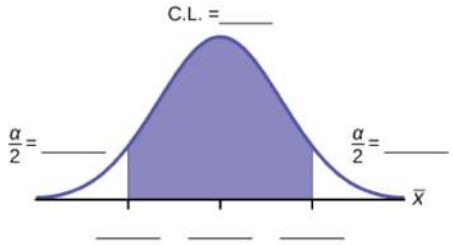


Figure 6.8.3

92. In one complete sentence, explain what the interval means.

93. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

94. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

95. Find the value of the sample size needed to if the confidence interval is 90% that the sample proportion and the population proportion are within 4% of each other. The sample proportion is 0.60. Note: Round all fractions up for n.

96. Find the value of the sample size needed to if the confidence interval is 95% that the sample proportion and the population proportion are within 2% of each other. The sample proportion is 0.650. Note: Round all fractions up for n.

97. Find the value of the sample size needed to if the confidence interval is 95% that the sample proportion and the population proportion are within 5% of each other. The sample proportion is 0.70. Note: Round all fractions up for n.

98. Find the value of the sample size needed to if the confidence interval is 90% that the sample proportion and the population proportion are within 1% of each other. The sample proportion is 0.50. Note: Round all fractions up for n.

99. Find the value of the sample size needed to if the confidence interval is 95% that the sample proportion and the population proportion are within 2% of each other. The sample proportion is 0.65. Note: Round all fractions up for n.

100. Find the value of the sample size needed to if the confidence interval is 95% that the sample proportion and the population proportion are within 4% of each other. The sample proportion is 0.45. Note: Round all fractions up for n.

101. Find the value of the sample size needed to if the confidence interval is 90% that the sample proportion and the population proportion are within 2% of each other. The sample proportion is 0.3. Note: Round all fractions up for n.

6.3 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

102. 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 95% confidence interval for the population proportion of Bam-Bam snack

pieces.

1. Define the random variables X and P' in words.
2. Which distribution should you use for this problem? Explain your choice
3. Calculate P'.
4. Construct a 95% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
5. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

103. 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.

1. - $\overline{x}$ = _____
    - s_x = _____
    - n = _____
    - $n - 1$ = _____
2. Define the random variables X and $\overline{x}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
    - State the confidence interval.
    - Sketch the graph.
5. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

104. Suppose that a committee is studying whether or not there is waste of time 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.

1. - $\overline{x}$ = _____
    - s_x = _____
    - n = _____
    - $n - 1$ = _____
2. Define the random variables X and $\overline{x}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population mean time wasted.
    - State the confidence interval.
    - Sketch the graph.
5. Explain in a complete sentence what the confidence interval means.

105. 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.

1. - $\overline{x}$ = _____
    - s_x = _____
    - n = _____
    - $n - 1$ = _____
2. Define the random variable X in words.
3. Define the random variable $\overline{x}$ in words.
4. Which distribution should you use for this problem? Explain your choice.

5. Construct a 95% confidence interval for the population mean length of time.
    - State the confidence interval.
    - Sketch the graph.
6. What does it mean to be “95% confident” in this problem?

106. 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.

1. - $\overline{x}$ = _____
    - s_x = _____
    - n = _____
    - $n - 1$ = _____
2. Define the random variable X in words.
3. Define the random variable $\overline{x}$ in words.
4. Which distribution should you use for this problem? Explain your choice.
5. Construct a 99% confidence interval for the population mean length of time using training wheels.
    - State the confidence interval.
    - Sketch the graph.
6. Why would the error bound change if the confidence level were lowered to 90%?

107. 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.

The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 30 Leadership PACs.

$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

Table 6.8.2

$\overline{x} = \$251,854.23$

$s = \$521,130.41$

Use this sample data to construct a 95% 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.

108. *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 table here shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Table 6.8.3

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms. Use the Student's *t*-distribution.

109. 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.

1. - $\overline{x}$ = _____
   - s_x = _____
   - n = _____
   - $n - 1$ = _____
2. Define the random variables X and $\overline{x}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 90% confidence interval for the population mean number of unoccupied seats per flight.
   - State the confidence interval.
   - Sketch the graph.

110. 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.

1. Which distribution should you use for this problem? Explain your choice.
2. Define the random variable $\overline{x}$ in words.
3. Construct a 95% confidence interval for the population mean cost of a used car.
   - State the confidence interval.
   - Sketch the graph.
4. Explain what a "95% confidence interval" means for this study.

111. Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

1. 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.
2. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
3. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.

4. Calculate the mean.
5. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

112. A survey of the mean number of cents off that coupons give 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¢. Assume the underlying distribution is approximately normal.

1. - $\overline{x}$ = _____
   - s_x = _____
   - n = _____
   - $n - 1$ = _____
2. Define the random variables X and $\overline{x}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population mean worth of coupons.
   - State the confidence interval.
   - Sketch the graph.
5. 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.

Use the following information to answer the next two exercises: 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.

113. Find the 95% confidence interval for the true population mean for the amount of soda served.

a. [12.42, 14.18]
b. [12.32, 14.29]
c. [12.50, 14.10]
d. Impossible to determine

6.4 A Confidence Interval for A Population Proportion

114. Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

1. 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?
2. 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?

115. 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.

1. - x = _____
   - n = _____
   - P' = _____
2. Define the random variables X and P', in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
   - State the confidence interval.
   - Sketch the graph.
5. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

116. According to a recent 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.

1. Define the random variables X and P' in words.

2. Which distribution should you use for this problem? Explain your choice.
3. 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.

117. An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latino, 323 identified themselves as Black, 254 identified themselves as Asian, and 779 identified themselves as White. In this survey, 86% of Black respondents said that they would welcome a White person into their families. Among Asian respondents, 77% would welcome a white person into their families, 71% would welcome a Latino person, and 66% would welcome a Black person.

1. 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 P', in words.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval.
    - State the confidence interval.
    - Sketch the graph.

118. Refer to the information in the prior question.

1. Construct three 95% confidence intervals.
    - percent of all Asians who would welcome a White person into their families.
    - percent of all Asians who would welcome a Latino into their families.
    - percent of all Asians who would welcome a Black person into their families.
2. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
3. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
4. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

119. 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.

1. Define the random variables X and P' in words.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% 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.
4. Explain what a “95% confidence interval” means for this study.

120. 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.” We are interested in the population proportion of adult Americans who feel that crime is the main problem.

1. Define the random variables X and P' in words.
2. Which distribution should you use for this problem? Explain your choice.
3. 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.
4. Suppose we want to lower the sampling error. What is one way to accomplish that?

5. 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.

121. Refer to the previous question. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three 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.

1. Define the random variables X and P' in words.
2. Which distribution should you use for this problem? Explain your choice.
3. 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.
4. 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.

Use the following information to answer the next three exercises: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. 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.

122. The point estimate for the true population proportion is:

a. 0.90
b. 1.27
c. 0.79
d. 400

123. A 90% confidence interval for the population proportion is _____.

a. [0.761, 0.820]
b. [0.125, 0.188]
c. [0.755, 0.826]
d. [0.130, 0.183]

Use the following information to answer the next two exercises: Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

124. 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.

a. [0.2975, 0.3796]
b. [0.6270, 0.6959]
c. [0.3041, 0.3730]
d. [0.6204, 0.7025]

125. The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

a. 0.6614
b. 0.3386
c. 173
d. 338

126. On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. 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.

1. Determine the estimated proportion from the sample.
2. Determine the sample size.
3. Identify CL and α.
4. Calculate the error bound based on the information provided.
5. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
6. Create a confidence interval for the results of this study.
7. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

127. A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 57% of Americans believe that big-time college sports programs corrupt the process of higher education.

1. Find the point estimate and the error bound for this confidence interval using 95% confidence.
2. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
3. Use the point estimate from part a and $n = 1,000$ to calculate a 90% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
4. Can we (with 90% confidence) conclude that at least half of all American adults believe this?

128. Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

1. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
2. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound 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.
3. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

129. 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 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

130. 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.

1. - $\overline{x}$ =_____
   - σ =_____
   - n =_____
2. In words, define the random variables X and $\overline{x}$.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population mean height of male Swedes.
   - State the confidence interval.
   - Sketch the graph.
5. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

131. 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.

1. In words, define the random variables X and $\overline{x}$.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval for the population mean length of engineering conferences.
   - State the confidence interval.
   - Sketch the graph.

132. 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.

1. - $\overline{x}$ =_____
   - σ =_____
   - n =_____
2. In words, define the random variables X and $\overline{x}$.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 90% confidence interval for the population mean time to complete the tax forms.
5. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
6. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
7. Suppose that the firm decided that it needed to be at least 95% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

133. 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.

1. - $\overline{x}$ =_____
   - σ =_____
   - s_x =_____
2. In words, define the random variable X.
3. In words, define the random variable $\overline{x}$.
4. Which distribution should you use for this problem? Explain your choice.
5. Construct a 90% confidence interval for the population mean weight of the candies.
   - State the confidence interval.
   - Sketch the graph.
6. Construct a 99% confidence interval for the population mean weight of the candies.
   - State the confidence interval.
   - Sketch the graph.
   - Calculate the error bound.
7. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.
8. In complete sentences, give an interpretation of what the interval in part f means.

134. A camp director is interested in the mean number of letters each child sends during his or her 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.

1. - $\overline{x}$ =_____
   - σ =_____
   - n =_____
2. Define the random variables X and $\overline{x}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 90% confidence interval for the population mean number of letters campers send home.
   - State the confidence interval.
   - Sketch the graph.
5. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

135. Which of the following options best matches what is meant by the term "90% confident" when constructing a confidence interval for a mean?

1. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
2. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
3. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
4. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

136. 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 table 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.

$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

Table 6.8.4

1. Find the point estimate for the population mean.
2. Using 95% confidence, calculate the error bound.
3. Create a 95% confidence interval for the mean total individual contributions.
4. Interpret the confidence interval in the context of the problem.

137. 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 U.S. falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

138. 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 college or university to within one inch with 95% confidence. How many male students must you measure?

139. If the confidence interval is change to a higher probability, would this cause a lower, or a higher, minimum sample size?

140. If the tolerance is reduced by half, how would this affect the minimum sample size?

141. If the value of P is reduced, would this necessarily reduce the sample size needed?

142. A company has been running an assembly line with 97.42% of the products made being acceptable. Then, a critical piece broke down. After the repairs the decision was made to see if the number of defective products made was still close enough to the long standing production quality. Samples of 500 pieces were selected at random, and the defective rate was found to be 0.025%.

1. Is this sample size adequate to claim the company is checking within the 90% confidence interval?
2. The 95% confidence interval?

6.8: Chapter 6 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.9: Chapter 6 Solutions

2. X is the number of hours a patient waits in the emergency room before being called back to be examined. $\overline{x}$ is the mean wait time of 70 patients in the emergency room.

4. CI: [1.3808, 1.6192], $EBM = 0.12$

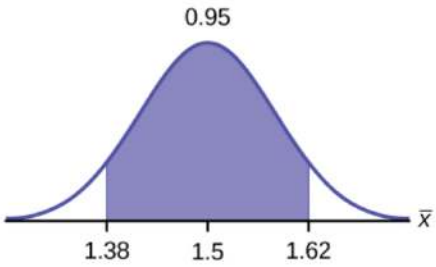


Figure 6.9.1

6.

1. $\overline{x} = 151$
2. $s_x = 32.6$
3. $n = 108$
4. $n - 1 = 107$

8. $\overline{x}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

10. CI: [142.92, 159.08], $EBM = 8.08$

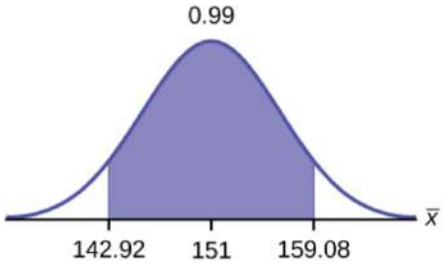


Figure 6.9.2

12.

1. 3.26
2. 1.02
3. 39

14. μ

16. t_{38}

18. 0.025

20. [2.93, 3.59]

22. We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

23. The error bound would become $EBM = 0.245$. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

26. It would decrease, because the *z*-score would decrease, which reducing the numerator and lowering the number.

28. X is the number of "successes" where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

30. CI: [0.5321, 0.6679], EBM : 0.0679

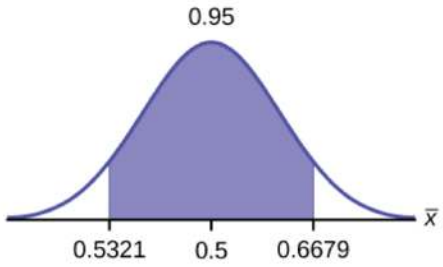


Figure 6.9.3

32. X is the number of "successes" where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

34. CI: [0.19432, 0.33068]

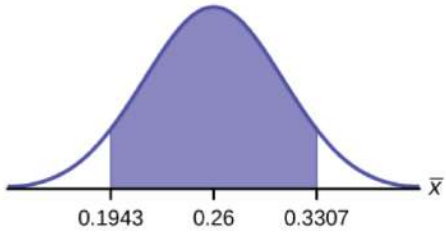


Figure 6.9.4

36. The sampling error means that the true mean can be 2% above or below the sample mean.

38. P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

40. CI: [0.62735, 0.67265], EBM: 0.02265

42. The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

44.

1. $x = 64$
2. $n = 100$
3. $P' = 0.64$

46. P

48. [0.55, 0.73]

50. 0.025

52. [0.55; 0.73]

54. With 95% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 55% and 73%.

56. The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

58.

1. 244

2. 15
3. 50

60. $N\left(244, \frac{15}{\sqrt{50}}\right)$

62. As the sample size increases, there will be less variability in the mean, so the interval size decreases.

64. X is the time in minutes it takes to complete the U.S. Census short form. $\overline{x}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

66. CI: [7.9441, 8.4559]

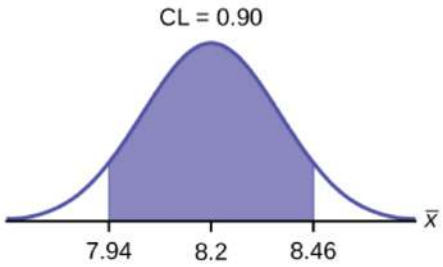


Figure 6.9.5

68. The level of confidence would decrease because decreasing *n* makes the confidence interval wider, so at the same error bound, the confidence level decreases.

70.

1. $\overline{x} = 2.2$
2. $\sigma = 0.2$
3. $n = 20$

72. $\overline{x}$ is the mean weight of a sample of 20 heads of lettuce.

74. $EBM = 0.07$, CI: [2.1264, 2.2736]

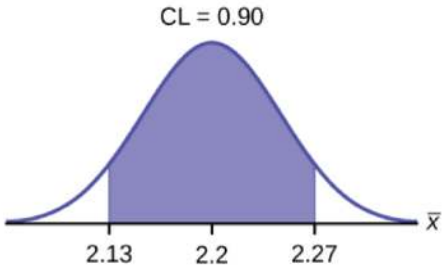


Figure 6.9.6

76. The interval is wider because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

78. The confidence level would increase.

80. 30.4

82. σ

84. μ

86. normal

88. 0.025

90. [24.52, 36.28]

92. We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

94. The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

96. 2,185

98. 6,765

100. 595

103.

1. - 8629
   - 6944
   - 35
   - 34
2. t_{34}
3. - CI: [6,244, 11,014]
   -

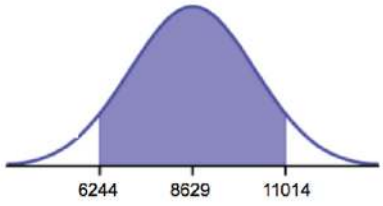


Figure 6.9.7

4. It will become smaller

105.

1. - $\overline{x} = 2.51$
   - $s_x = 0.318$
   - $n = 9$
   - $n - 1 = 8$
2. the effective length of time for a tranquilizer
3. the mean effective length of time of tranquilizers from a sample of nine patients
4. We need to use a Student's *t*-distribution, because we do not know the population standard deviation.
5. - CI: [2.27, 2.76]
   - Check student's solution.
6. If we were to sample many groups of nine patients, 95% of the CIs constructed from these samples would contain the true population mean length of time.

107. [$57,282.79, $446,425.67]

109.

1. - $\overline{x} = 11.6$
   - $s_x = 4.1$
   - $n = 225$
   - $n - 1 = 224$
2. X is the number of unoccupied seats on a single flight. $\overline{x}$ is the mean number of unoccupied seats from a sample of 225 flights.
3. We will use a Student's *t*-distribution, because we do not know the population standard deviation.
4. - CI: [11.15 , 12.05]

- Check student's solution.

111.

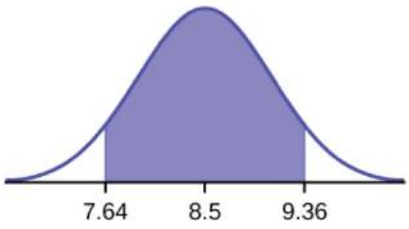
1. - CI: [7.64 , 9.36]
   - 

   Figure 6.9.8
2. The sample should have been increased.
3. Answers will vary.
4. Answers will vary.
5. Answers will vary.

113. b

114.

1. 1,068
2. The sample size would need to be increased since the critical value increases as the confidence level increases.

116.

1. X = the number of people who feel that the president is doing an acceptable job;

   P' = the proportion of people in a sample who feel that the president is doing an acceptable job.
2. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$
3. - CI: [0.59, 0.63]
   - Check student's solution

118.

1. - [0.72, 0.82]
   - [0.65, 0.76]
   - [0.60, 0.72]
2. Yes, the intervals [0.72, 0.82] and [0.65, 0.76] overlap, and the intervals [0.65, 0.76] and [0.60, 0.72] overlap.
3. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
4. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

120.

1. X = the number of adult Americans who feel that crime is the main problem; P^{prime} = the proportion of adult Americans who feel that crime is the main problem
2. Since we are estimating a proportion, given $P^{prime} = 0.2$ and $n = 1000$, the distribution we should use is $N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.
3. - CI: [0.18, 0.22]
   - Check student's solution.
4. One way to lower the sampling error is to increase the sample size.

5. The stated "±3" represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

122. c

125. a

127.

1. $P' = \frac{(0.57+0.49)}{2} = 0.53; EBP = 0.57 - 0.53 = 0.04$
2. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.
3. [0.50, 0.56]
4. Yes, the confidence interval no longer includes values less than 0.50.

130.

1. - 71
   - 2.8
   - 48
2. X is the height of a male Swede, and $\overline{x}$ is the mean height from a sample of 48 male Swedes.
3. Normal. We know the standard deviation for the population, and the sample size is greater than 30.
4. - CI: [70.151, 71.85]
   -

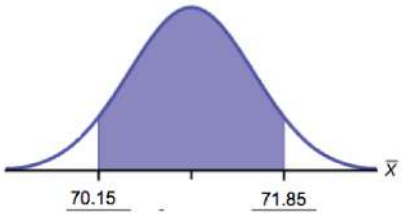


Figure 6.9.9

5. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

132.

1. - $\overline{x} = 23.6$
   - $\sigma = 7$
   - $n = 100$
2. X is the time needed to complete an individual tax form. $\overline{x}$ is the mean time to complete tax forms from a sample of 100 customers.
3. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.
4. [22.45, 24.75]
5. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.
6. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.
7. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

134.

1. - 7.9
   - 2.5

- 20

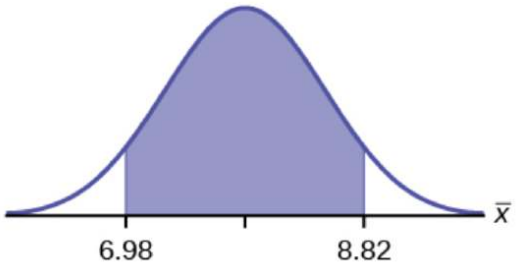
2. X is the number of letters a single camper will send home. $\overline{x}$ is the mean number of letters sent home from a sample of 20 campers.
3. $N\left(7.9, \frac{2.5}{\sqrt{20}}\right)$ because sigma is known.
4. - CI: [6.98, 8.82]
   - 

   Figure 6.9.10
5. The error bound and confidence interval will decrease because the standard error will get smaller.

136.

1. $\overline{x} = \$568,873$
2. $CL = 0.95\ \alpha = 1–0.95 = 0.05\ z_{\frac{\alpha}{2}} = 1.96$
   $EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\frac{909200}{\sqrt{40}} = \$281,764$
3. $\overline{x} - EBM = 568,873 - 281,764 = 287,109$
   $\overline{x} + EBM = 568,873 + 281,764 = 850,637$
4. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between $287,109 and $850,637.

139. Higher

140. It would increase to four times the prior value.

141. No, it could have no effect if it were to change to $1–P$, for example. If it gets closer to 0.5 the minimum sample size would increase.

142.

1. No
2. No

6.9: Chapter 6 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.10: Chapter 6 References

6.2 A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size

“American Fact Finder.” U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...html?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...phicTrends.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/2000...thchart-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/...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/disclosur...esummary.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).

6.3 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

“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 http://www.businessweek.com/.

Data from 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...tero%2Fnewborn (accessed July 2, 2013).

“Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at http://www.fec.gov/finance/disclosur...pPacList.shtml (accessed July 2, 2013).

6.4 A Confidence Interval for A Population Proportion

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/2...d-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//...al%20Media.pdf (accessed July 2, 2013).

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/th...ement-age.aspx (accessed July 2, 2013).

The Field Poll. Available online at http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).

Zogby. “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/2...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/publ...cation_process (accessed July 2, 2013).

6.10: Chapter 6 References is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

7: Hypothesis Testing with One Sample

7.1: Introduction to Hypothesis Testing
7.2: Null and Alternative Hypotheses
7.3: Outcomes and Type I and Type II Errors
7.4: Distribution Needed for Hypothesis Testing
7.5: Full Hypothesis Test Examples
7.6: Chapter 7 Key Terms
7.7: Chapter 7 Review
7.8: Chapter 7 Formula Review
7.9: Chapter 7 Homework
7.10: Chapter 7 Solutions
7.11: Chapter 7 References

This page titled 7: Hypothesis Testing with One Sample is shared under a CC BY license and was authored, remixed, and/or curated by .

7.1: Introduction to Hypothesis Testing

Now we are down to the bread and butter work of the statistician: developing and testing hypotheses. It is important to put this material in a broader context so that the method by which a hypothesis is formed is understood completely. Using textbook examples often clouds the real source of statistical hypotheses.

Statistical testing is part of a much larger process known as the scientific method. This method was developed more than two centuries ago as the accepted way that new knowledge could be created. Until then, and unfortunately even today, among some, "knowledge" could be created simply by some authority saying something was so, *ipso dicta*. Superstition and conspiracy theories were (are?) accepted uncritically.



Figure 7.1.1 You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

The scientific method, briefly, states that only by following a careful and specific process can some assertion be included in the accepted body of knowledge. This process begins with a set of assumptions upon which a theory, sometimes called a model, is built. This theory, if it has any validity, will lead to predictions; what we call hypotheses.

As an example, in microeconomics the theory of consumer choice begins with certain assumptions concerning human behavior. From these assumptions followed a theory of how consumers make choices using indifference curves and the budget line. This theory gave rise to a very important prediction; namely, that there was an inverse relationship between price and quantity demanded. This relationship was known as the demand curve. The negative slope of the demand curve is really just a prediction, or a hypothesis, that can be tested with statistical tools.

Unless hundreds and hundreds of statistical tests of this hypothesis had not confirmed this relationship, the so-called Law of Demand would have been discarded years ago. This is the role of statistics, to test the hypotheses of various theories to determine if they should be admitted into the accepted body of knowledge, and how we understand our world. Once admitted, however, they may be later discarded if new theories come along that make better predictions.

Not long ago two scientists claimed that they could get more energy out of a process than was put in. This caused a tremendous stir for obvious reasons. They were on the cover of *Time* and were offered extravagant sums to bring their research work to private industry and any number of universities. It was not long until their work was subjected to the rigorous tests of the scientific method and found to be a failure. No other lab could replicate their findings. Consequently they have sunk into obscurity and their theory discarded. It may surface again when someone can pass the tests of the hypotheses required by the scientific method, but until then it is just a curiosity. Many pure frauds have been attempted over time, but most have been found out by applying the process of the scientific method.

This discussion is meant to show just where in this process statistics falls. Statistics and statisticians are not necessarily in the business of developing theories, but in the business of testing others' theories. Hypotheses come from these theories based upon an explicit set of assumptions and sound logic. The hypothesis comes first, before any data are gathered. Data do not create hypotheses; they are used to test them. If we bear this in mind as we study this section the process of forming and testing hypotheses will make more sense.

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 the value of a specific parameter. For instance, a car dealer 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 women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called "**hypothesis testing**". 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, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

This page titled 7.1: Introduction to Hypothesis Testing is shared under a CC BY license and was authored, remixed, and/or curated by .

7.2: Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative (or research) hypothesis**. These hypotheses contain opposing viewpoints.

- H_a: **The alternative (or research) hypothesis:** This is a claim about the predicted effect in a population. It is contradictory to H_0 and what we conclude when we reject H_0. For example, we might like to test whether the mean of men's salaries and the mean of women's salaries in a given occupation differ from each other.
- H_0: **The null hypothesis:** This is a statement of no difference (that is, zero difference) between a sample mean or proportion and a population mean or proportion. This is used in hypothesis testing since it allows us to set a clear baseline for comparison. For example, we can compare whether there is more than 0 difference between men's mean salaries and women's mean salaries in a given occupation. When conducting hypothesis testing, if you reject the null it suggests there is an effect.

Since the null hypothesis provides a fixed baseline for testing the alternative hypothesis, 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 whether the sample data appear likely under the null hypothesis, you make a **decision.** There are two options for a decision, and you always make the decision about the null hypothesis. The two decision options are (1) "reject H_0" if the sample data seems unlikely to have occurred according to the null hypothesis; or (2) "do not reject H_0" or "decline to reject H_0" or "fail to reject H_0" if the sample information seems probable within the prediction made by the null hypothesis. These conclusions are all based upon a set level of probability, a significance level, that is chosen by the analyst. Typically, the selected probability level is 5% or 1%, but we'll discuss what helps decide the exact probability here later.

Table 7.2.1 presents possible forms of hypotheses in pairs. For example, if the null hypothesis is equal to some value, the alternative has to be not equal to that value.

Table 7.2.1

H_0	H_a
equal (=)	not equal ($\neq$)
greater than or equal to ($\geq$)	less than (<)
less than or equal to ($\leq$)	more than (>)

Note

As a mathematical convention 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 depends on the wording of the hypothesis test.

Example 7.2.1

H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $P \leq .30$
H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $P > .30$

Example 7.2.2

We want to test whether the mean GPA of students in American colleges is different from 3.0 (out of 4.0). The null and alternative hypotheses are:
$H_0 : \mu = 3.0$
$H_a : \mu \neq 3.0$

Example 7.2.3

We want to test if college students take less than five years to graduate from college, on average. The null and alternative hypotheses are:

$H_0 : \mu \geq 5$

$H_a : \mu < 5$

This page titled 7.2: Null and Alternative Hypotheses is shared under a CC BY license and was authored, remixed, and/or curated by .

7.3: Outcomes and Type I and Type II Errors

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:

Table 7.3.1

Statistical Decision	H_0 is actually...	
	True	**False**
Reject H_0	Type I error	Correct outcome
Do not reject H_0	Correct outcome	Type II error

The four possible outcomes in the table are:

1. The decision is **reject** H_0 when H_0 **is true** (incorrect decision known as a **Type I error**). This can be thought of as a "false positive" or "false alarm". As we will see later, it is this type of error that we will guard against by setting a small probability of making such an error.
2. The decision is **reject** H_0 when H_0 **is false** (**correct decision**).
3. The decision is **do not reject** H_0 when H_0 **is true (correct decision).**
4. The decision is **do not reject** H_0 when, in fact, H_0 **is false** (incorrect decision known as a **Type II error**). This can be thought of as a "false negative" or "miss".

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: rejecting a good null.
- β = probability of a Type II error = P**(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false. $(1-\beta)$ is called the **power of the test**, or **statistical power**.

α and β should be as small as possible because they are probabilities of errors.

Statistics allows us to set the probability that we are making a Type I error. The probability of making a Type I error is α. Recall that the confidence intervals in the last unit were set by choosing a value called z_α (or t_α) and the alpha value determined the confidence level of the estimate because it was the probability of the interval failing to capture the true mean (or true proportion P). This alpha and that one are the same.

The easiest way to see the relationship between the alpha error and the level of confidence is with the following figure.

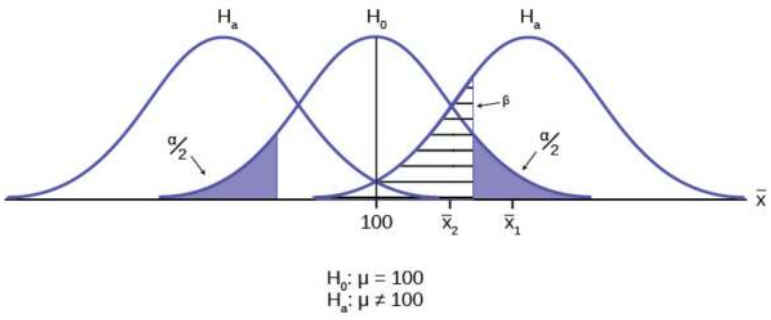


Figure 7.3.1

In the center of Figure 7.3.1 is a normally distributed sampling distribution marked H_0. This is a sampling distribution of $\overline{x}$ and by the Central Limit Theorem it is normally distributed. The distribution in the center is marked H_0 and represents the distribution for the null hypotheses H_0: $\mu = 100$. This is the value that is being tested. The formal statements of the null and alternative hypotheses are listed below the figure.

The distributions on either side of the H_0 distribution represent distributions that would be true if H_0 is false, under the alternative hypothesis listed as H_a. We do not know which is true, and will never know. There are, in fact, an infinite number of distributions from which the data could have been drawn if H_a is true, but only two of them are on Figure 7.3.1 representing all of the others.

To test a hypothesis we take a sample from the population and determine if it could have come from the hypothesized distribution with an acceptable level of significance. This level of significance is the alpha error and is marked on Figure 7.3.1 as the shaded areas in each tail of the H_0 distribution. (Each area is actually $\alpha/2$ because the distribution is symmetrical and the alternative hypothesis in this example allows for the possibility for the value to be either greater than or less than the hypothesized value--called a **two-tailed test**).

If the sample mean marked as $\overline{x}_1$ is in the tail of the distribution of H_0, we conclude that the probability that it could have come from the H_0 distribution is less than alpha. We consequently state, "the null hypothesis is rejected at α level of significance". The truth **may** be that this $\overline{x}_1$ did come from the H_0 distribution, but from out in the tail. If this is so then we have erroneously rejected a true null hypothesis and have made a Type I error. What statistics has done is provide an estimate about what we know, and what we control, and that is the probability of us being wrong, α.

We can also see in Figure 7.3.1 that the sample mean could be really from an H_a distribution, but within the boundary set by the alpha level. Such a case is marked as $\overline{x}_2$. There is a probability that $\overline{x}_2$ actually came from H_a but shows up in the range of H_0 between the two tails. This probability is the beta error, the probability of failing to reject a false null.

Our problem is that we can only set the alpha error because there are an infinite number of alternative distributions from which the mean could have come that are not equal to H_0. As a result, the statistician places the burden of proof on the alternative hypothesis. That is, we will not reject a null hypothesis unless there is a greater than 90, or 95, or even 99 percent probability that the null is false: the burden of proof lies with the alternative hypothesis. This is why we call this the tyranny of the status quo.

By way of example, the American judicial system begins with the concept that a defendant is "presumed innocent". This is the status quo and is the null hypothesis. The judge will tell the jury that they can not find the defendant guilty unless the evidence indicates guilt beyond a "reasonable doubt" which is usually defined in criminal cases as 95% certainty of guilt. If the jury rejects the null, innocence, then action will be taken, jail time. The null hypothesis is the "default", so the burden of proof always lies with the alternative hypothesis. (In civil cases, the jury needs only to be more than 50% certain of wrongdoing to find culpability, called "a preponderance of the evidence").

The example above was for a test of a single mean, but the same logic applies to tests of hypotheses for all statistical parameters one may wish to test.

The following are examples of Type I and Type II errors.

Example 7.3.1

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.

Type I error: Frank thinks that his rock climbing equipment is not safe (that is, rejecting the null) when, in fact, it really is safe (that is, the null is really true).

Type II error: Frank thinks that his rock climbing equipment is safe (that is, failing to reject the null) when, in fact, it is not safe (that is, the null is really false).

$\alpha =$ **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe.

$\beta =$ **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. (Frank will be using unsafe climbing equipment!)

Example 7.3.2

Suppose the null hypothesis, H_0, is: A particular work-from-home training program helps remote workers be more productive, as shown by an employee survey after the training.

Type I error: The employee survey shows that the training program is effective when, in fact, the training program is not effective.

Type II error: The employee survey shows that the training program is not effective when, in fact, the training program is effective.

$\alpha =$ **probability** that the employee survey shows that the training program is effective when, in fact, the training program is not effective = P(Type I error).

$\beta =$ **probability** that the employee survey shows that the training program is not effective when, in fact, the training program is effective = P(Type II error).

In this case, one could argue that both a Type I error and a Type II error could have important consequences for the company. If we commit a Type I error, we could be using an ineffective training program, which may waste company and worker resources. On the other hand, if we commit a Type II error, we could decide to cut the program, thinking it is not working, but we would be missing out on a truly effective training program for our workers.

Exercise 7.3.1

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?

Example 7.3.3

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 error: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.

Type II error: 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%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

This page titled 7.3: Outcomes and Type I and Type II Errors is shared under a CC BY license and was authored, remixed, and/or curated by .

7.4: Distribution Needed for Hypothesis Testing

Earlier, we discussed sampling distributions. Particular distributions are associated with hypothesis testing. We will perform hypotheses tests of a population mean using either a normal distribution or a Student's t-distribution. (Remember, use a Student's t-distribution when the population standard deviation is unknown and the sample size is small, where small is considered to be less than 100 observations.) We perform tests of a population proportion using a normal distribution when we can assume that the distribution is normally distributed. We consider this to be true if the sample size is 100 or more. This is the same rule of thumb we used when developing the formula for the confidence interval for a population proportion.

Hypothesis Test for the Mean

Going back to the standardizing formula we can derive the **test statistic** for testing hypotheses concerning means.

$$z_{obs} = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

The standardizing formula can not be solved as it is because we do not have μ, the population mean. However, if we substitute in the hypothesized value of the mean, μ_0 in the formula as above, we can compute a z value. This is the test statistic for a test of hypothesis for a mean and is presented in Figure 7.4.1. We interpret this z value as the associated probability that a sample with a sample mean of $\overline{x}$ could have come from a distribution with a population mean of H_0 and we call this z value z_{obs} for “observed”. At times notation z_c for “calculated” is used (see figures below). Figure 7.4.1 and Figure 7.4.2 show this process.

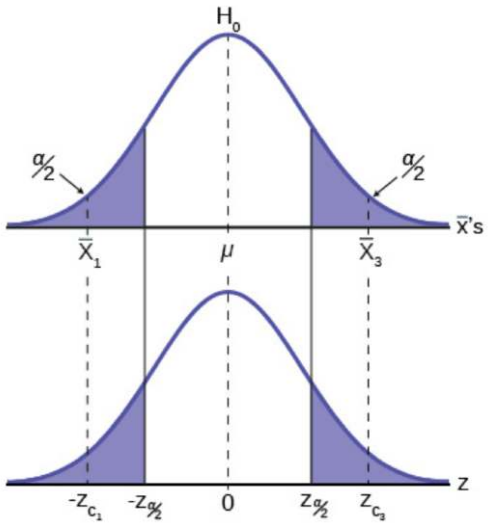


Figure 7.4.1

In Figure 7.4.1 two of the three possible outcomes are presented. $\overline{x}_1$ and $\overline{x}_3$ are in the tails of the hypothesized distribution of H_0. Notice that the horizontal axis in the top panel is labeled $\overline{x}$'s. This is the same theoretical distribution of $\overline{x}$'s, the sampling distribution, that the Central Limit Theorem tells us is normally distributed. This is why we can draw it with this shape. The horizontal axis of the bottom panel is labeled z and is the standard normal distribution. $z_{\frac{\alpha}{2}}$ and $-z_{\frac{\alpha}{2}}$, called the **critical values**, are marked on the bottom panel as the z values associated with the probability the analyst has set as the level of significance in the test, α. The probabilities in the tails of both panels are, therefore, the same.

Notice that for each $\overline{x}$ there is an associated z or z_c, called the observed or calculated z, that comes from solving the equation above. This observed z is nothing more than the number of standard errors that the sample mean is from the **hypothesized** mean. If the sample mean falls "too many" standard errors from the hypothesized mean we conclude that the **sample** mean could not have come from the distribution with the hypothesized mean, given our pre-set required level of significance. It **could** have come from H_0, but it is deemed just too unlikely. In Figure 7.4.1 both $\overline{x}_1$ and $\overline{x}_3$ are in the tails of the distribution. They are deemed "too far" from the hypothesized value of the mean given the chosen level of alpha. If in fact this sample mean it did come from H_0, but from in the tail, we have made a Type I error: we have rejected a good null. Our only real comfort is that we know the probability of making such an error, α, and we can control the size of α.

Figure 7.4.2 shows the third possibility for the location of the sample mean, $\overline{x}$. Here the sample mean is within the two critical values. That is, within the probability of 1 - α and we cannot reject the null hypothesis.

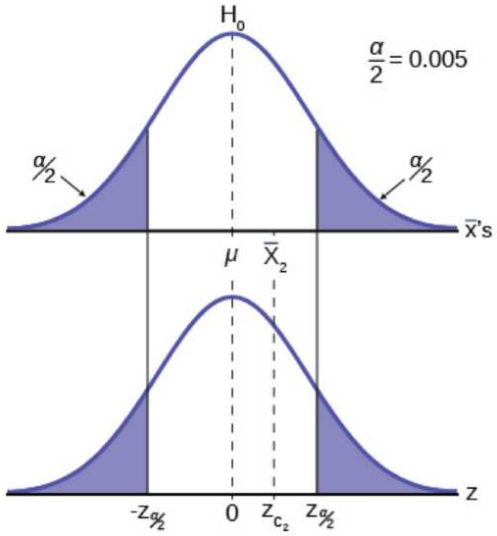


Figure 7.4.2

This gives us the decision rule for testing a hypothesis for a two-tailed test:

Decision rule: two-tailed test
If $\lvert z_{obs}\rvert \leq z_{\frac{\alpha}{2}}$: then do **not** reject H_0
If $\lvert z_{obs}\rvert > z_{\frac{\alpha}{2}}$: then reject H_0

Table 7.4.1

This rule will always be the same no matter what hypothesis we are testing or what formulas we are using to make the test. The only change will be to change the z_{obs} to the appropriate symbol for the test statistic for the parameter being tested. Stating the decision rule another way: if the sample mean is unlikely to have come from the distribution with the hypothesized mean we must reject the null hypothesis. Here we define "unlikely" as having a probability less than alpha of occurring.

p-value Approach

An alternative decision rule can be developed by calculating the probability of observing the sample mean $\overline{x}$ if we **assume the null hypothesis is true**. This probability would, in other words, equal the tail probability that the t-test statistic is equal or greater than our observed test statistic z_{obs} from sample data (under the assumption that H_0 is true).

Here the notion of "likely" and "unlikely" is defined by the probability of drawing a sample with a mean $\overline{x}$ from a population with the hypothesized mean that is either larger or smaller than that found in the sample data. Simply stated, the p-value approach compares the desired significance level, α, to the p-value which is the probability of drawing a sample mean further from the hypothesized value than the actual sample mean, assuming H_0 is true. A large p-value calculated from the data indicates this is a likely outcome and 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. The relationship between the decision rule of comparing the observed (calculated) test statistic, z_{obs}, and the critical value, z_{α}, and using the p-value can be seen in Figure 7.4.3.

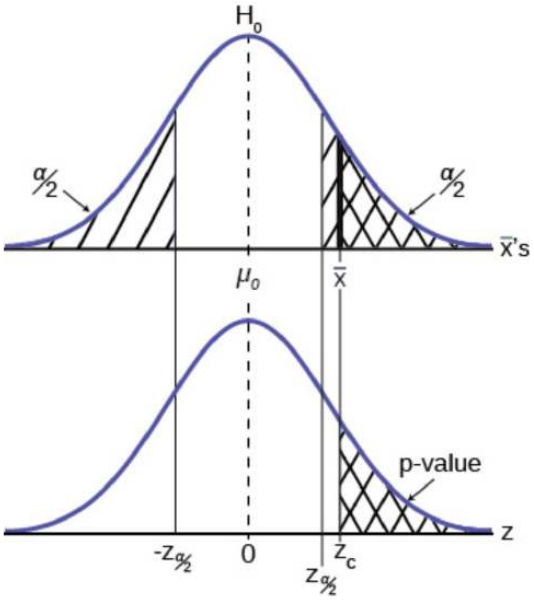


Figure 7.4.3

The observed (calculated) value of the test statistic is z_c in this figure and is marked on the bottom graph of the standard normal distribution because it is a z value. In this case the calculated value is in the tail and thus we must reject the null hypothesis. The associated $\overline{x}$ is just unusually large (and thus too unlikely) to believe that it came from the distribution with a mean of μ_0 with a significance level of α.

If we use the p-value decision rule we need one more step. We need to find in the standard normal table the probability associated with the observed test statistic, z_{obs}. We then compare that to the α associated with our selected level of confidence. In Figure 7.4.3 we see that the p-value is less than α and therefore we must reject the null. We know that the p-value is less than α because the area under the p-value is smaller than $\alpha/2$. When our sample size is small (having less than 100 observations), we can find an approximate p-value from the Student's t table. In the relevant df row based on our sample size, we first locate the test statistic value(s) closest to our observed t-statistic, t_{obs}. We then find the one-tailed or two-tailed probabilities associated with those test statistic value(s) depending on the type of test we are conducting (see next section). In case of a two-tailed test, the one-tailed probabilities from the Student's t table need to be multiplied by two. The obtained p-value interval will give us the range that we will then compare to α and make our decision.

It is important to note that two researchers drawing randomly from the same population may find two different p-values from their samples. This occurs because the p-value is calculated as the probability in the tail beyond the sample mean assuming that the null hypothesis is correct. Because the sample means will in all likelihood be different this will create two different p-values.

Here is a systematic way to make a decision of whether you reject or do not reject a null **hypothesis** if using the p-**value** and a **preset or preconceived** α (the "**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. In any case, the value of α is the decision of the analyst. When you make a decision to reject or not reject H_0, do as follows:

- If p-value < α, reject H_0. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If p-value $\geq \alpha$, cannot reject H_0. There is not sufficient evidence to conclude that the alternative hypothesis, H_a, may be correct. In this case the status quo - the baseline set by the null hypothesis - stands.
- When you "cannot 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_0; and we must remember that the null is only a baseline comparison point, set up for testing our alternative or research hypothesis, H_a.

One and Two-tailed Tests

The discussion of Figure 7.4.1 - Figure 7.4.3 was based on the null and alternative hypothesis presented in Figure 7.4.1. This was called a **two-tailed test** because the alternative hypothesis allowed that the mean could have come from a population which was **either** larger or smaller than the hypothesized mean in the null hypothesis. This could be seen by the statement of the alternative hypothesis as $\mu \neq 100$, in this example.

It may be that the analyst has an interest in whether the sample differs in a certain **direction** from (for example, is higher than) the hypothesized value. If this is the case, it becomes a **one-tailed test** and all of the alpha probability is placed in just one tail and not split into $\alpha/2$ as in the above case of a two-tailed test. For example, a car manufacturer claims that their Model 17B provides gas mileage of greater than 25 miles per gallon. The null and alternative hypothesis would be:

- $H_0 : \mu \leq 25$
- $H_a : \mu > 25$

The claim would be in the alternative hypothesis. The burden of proof in hypothesis testing is carried in the alternative. This is because failing to reject the null, the status quo, must be accomplished with 90 or 95 percent significance that it cannot be maintained. Said another way, we want to have only a 5 or 10 percent probability of making a Type I error, rejecting a good null; overthrowing the status quo.

Figure 7.4.4 shows the two possible cases and the form of the null and alternative hypothesis that give rise to them, where μ_0 is the hypothesized value of the population mean.

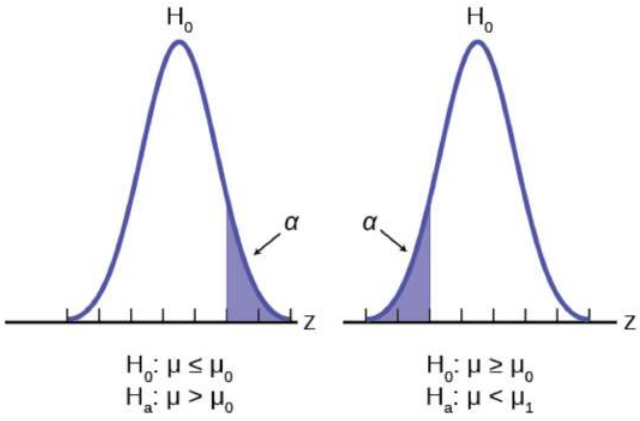


Figure 7.4.4

Effects of Sample Size on Test Statistic

In developing the confidence intervals for the mean from a sample, we found that most often we would not have the population standard deviation, σ. If the sample size were less than 100, we could simply substitute the point estimate for σ, the sample standard deviation, s, and use the Student's t-distribution to correct for this lack of information.

When testing hypotheses we are faced with this same problem and the solution is exactly the same. Namely, if the population standard deviation is unknown, and the sample size is less than 100, substitute s, the point estimate for the population standard deviation, σ, in the formula for the test statistic and use the Student's t distribution. All the formulas and figures above are unchanged except for this substitution and changing the Z distribution to the Student's t distribution on the graph. Remember that the Student's t distribution can only be computed knowing the proper degrees of freedom for the problem. In this case, the degrees of freedom is computed as before with confidence intervals: $df = (n - 1)$. The observed t-value is compared to the t-value associated with the pre-set level of confidence required in the test, $t_{\alpha,df}$ found in the Student's t tables. If we do not know σ, but the sample size is 100 or more, we simply substitute s for σ and use the normal distribution.

Table 7.4.2 summarizes these rules.

Sample size	Test statistic
< 100 (σ unknown)	$t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
< 100 (σ known)	$z_{obs} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
> 100 (σ unknown)	$z_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
> 100 (σ known)	$z_{obs} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

Table 7.4.2 Test Statistics for Test of Means, Varying Sample Size, Population Standard Deviation Known or Unknown

A Systematic Approach for Testing a Hypothesis

A systematic approach to hypothesis testing follows the following steps and in this order. This template will work for all hypotheses that you will ever test.

- Set up the null and alternative hypothesis. This is typically the hardest part of the process. Here the question being asked is reviewed. What parameter is being tested, a mean, a proportion, differences in means, etc. Is this a one-tailed test or two-tailed test?
- Decide the level of significance required for this particular case and determine the critical value. These can be found in the appropriate statistical table. The levels of confidence typical for businesses are 90, 95, and 99. However, the level of significance is a policy decision and should be based upon the risk of making a Type I error, rejecting a good null. Consider the consequences of making a Type I error.

  Next, on the basis of the hypotheses and sample size, select the appropriate test statistic and find the relevant critical value: z_α, t_α, etc. Drawing the relevant probability distribution and marking the critical value is always big help. Be sure to match the graph with the hypothesis, especially if it is a one-tailed test.
- Take a sample(s) and calculate the relevant parameters: sample mean, standard deviation, or proportion. Using the formula for the test statistic from above in step 2, now calculate the test statistic for this particular case using the parameters you have just calculated.
- Compare the calculated test statistic and the critical value. Marking these on the graph will give a good visual picture of the situation. There are now only two situations:
  1. The test statistic is in the tail: Reject the null, the probability that this sample mean (proportion) came from the hypothesized distribution is too small to believe that it is the real home of these sample data.
  2. The test statistic is not in the tail: Cannot reject the null, the sample data are compatible with the hypothesized population parameter.
- Reach a conclusion. It is best to articulate the conclusion two different ways. First a formal statistical conclusion such as “With a 5 % level of significance we must reject the null hypothesis that the population mean is equal to XX (units of measurement)”. The second statement of the conclusion is less formal and states the action, or lack of action, required. If the formal conclusion was that above, then the informal one might be, “The machine is broken and we need to shut it down and call for repairs”.

All hypotheses tested will go through this same process. The only changes are the relevant formulas and those are determined by the hypothesis required to answer the original question.

This page titled 7.4: Distribution Needed for Hypothesis Testing is shared under a CC BY license and was authored, remixed, and/or curated by .

7.5: Full Hypothesis Test Examples

Tests on Means

Example 7.5.1

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle.

His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds**, with a **standard deviation of 0.8 seconds**. **Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.** Conduct a hypothesis test using test statistics and p-values with a preset $\alpha = 0.05$.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

Set the null and alternative hypothesis:

In this case there is an implied challenge or claim. This is that the goggles will reduce the swimming time. The effect of this is to set the hypothesis as a one-tailed test. The claim will always be in the alternative hypothesis because the burden of proof always lies with the alternative. Remember that the status quo must be defeated with a high degree of confidence, in this case 95% confidence. The null and alternative hypotheses are thus:

$H_0 : \mu \geq 16.43 \qquad H_a : \mu < 16.43$

For Jeffrey to swim faster, his time should be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\overline{x}$ = the mean time to swim the 25-yard freestyle.

Distribution for the test statistic:

The sample size is less than 30 and we do not know the population standard deviation so this is a *t*-test. The proper formula is: $t_{obs} = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$

$\mu_0 = 16.43$ comes from H_0 and not the data. $\overline{x} = 16$, $s = 0.8$, and $n = 15$.

Our step 2, setting the level of confidence, has already been determined by the problem, α of .05 corresponds to a 95% confidence level. It is worth thinking about the meaning of this choice. The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds or more. (Reject the null hypothesis when the null hypothesis is true.) For this case the only concern with a Type I error would seem to be that Jeffrey's dad may fail to bet on his son's victory because he does not have appropriate confidence in the effect of the goggles.

To find the critical value we need to select the appropriate test statistic. We have concluded that this is a *t*-test on the basis of the sample size and that we are interested in a population mean. We can now draw the graph of the *t*-distribution and mark the critical value. For this problem the degrees of freedom are n-1, or 14. Looking up 14 degrees of freedom at the 0.05 column of the *t*-table we find 1.761. This is the critical value and we can put this on our graph.

Step 3 is the calculation of the test statistic using the formula we have selected. We find that the observed test statistic is -2.08, meaning that the sample mean is 2.08 standard errors below the hypothesized mean of 16.43.

$$t_{obs} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{16 - 16.43}{.8/\sqrt{15}} = -2.08$$

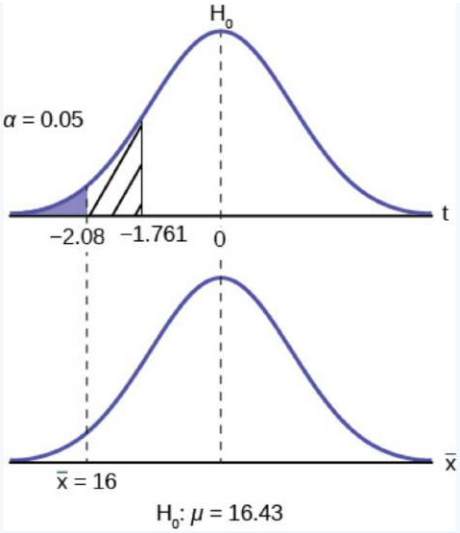


Figure 7.5.1

Step 4 has us compare the test statistic and the critical value and mark these on the graph. We see that the test statistic is in the tail and thus we move to step 4 and reach a conclusion. The probability that an average time of 16 minutes could come from a distribution with a population mean of 16.43 minutes is too unlikely to have occurred under the null hypothesis. We reject the null.

Step 5 has us state our conclusions first formally and then less formally. A formal conclusion would be stated as: “With a 95% level of confidence we reject the null hypothesis that the swimming time with goggles comes from a distribution with a population mean time of 16.43 minutes.” Less formally, “With 95% confidence, we believe that the goggles improved swimming speed".

If we wished to use the p-value system of reaching a conclusion we would calculate the statistic and take the additional step to find the probability of being 2.08 standard errors from the mean on a *t*-distribution. The p-value interval is (.025, .05), that we get by looking up the one-tailed probabilities associated with the closest *t*-scores (1.761 and 2.145) to the observed test statistic (-2.08) in the relevant *df* row of 14 in the *t*-table. Comparing this interval to the significance level of .05 we see that we reject the null. The p-value has been put on the graph as the shaded area beyond -2.08 and it shows that it is smaller than the hatched area which is the α level of 0.05. Both methods reach the same conclusion that we reject the null hypothesis.

Exercise 7.5.1

The mean throwing distance of a football for Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco’s mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Example 7.5.2

Jane has just begun her new job as on the sales force of a very competitive company. In a sample of 16 sales calls it was found that she closed the contract for an average value of 108 dollars with a standard deviation of 12 dollars. Company policy requires that new members of the sales force must exceed an average of $100 per contract during the trial employment period. Can we conclude that Jane has met this requirement at the significance level of 5%?

Answer

1. $H_0 : \mu \leq 100$
   $H_a : \mu > 100$
   The null and alternative hypothesis are for the parameter μ because the number of dollars of the contracts is a continuous random variable. Also, this is a one-tailed test because the company has only an interested if the number of dollars per contact is below a particular number not "too high" a number. This can be thought of as making a claim that the requirement is being met and thus the claim is in the alternative hypothesis.
2. Test statistic: $t_{obs} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{108-100}{\left(\frac{12}{\sqrt{16}}\right)} = 2.67$
3. Critical value: $t_\alpha = 1.753$ with $n - 1$ degrees of freedom = 15

The test statistic is a Student's *t* because the sample size is below 100; therefore, we cannot use the normal distribution. Comparing the observed value of the test statistic and the critical value of *t* at a 5% significance level, we see that the observed value is in the tail of the distribution. Thus, we conclude that 108 dollars per contract is significantly larger than the hypothesized value of 100 and thus we must reject the null hypothesis. There is evidence that Jane's performance meets company standards.

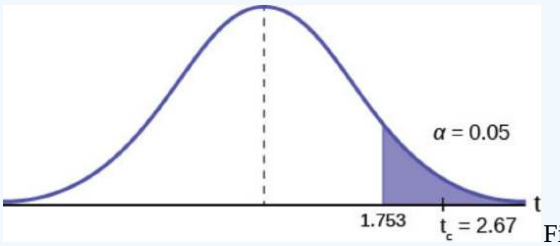


Figure 7.5.2

Exercise 7.5.2

It is believed that a stock price for a particular company will grow at a rate of $5 per week with a standard deviation of $1. An investor believes the stock won't grow as quickly. The changes in stock price is 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, state your conclusion, and identify the Type I and Type II errors.

Example 7.5.3

A manufacturer of salad dressings uses machines to dispense liquid ingredients into bottles that move along a filling line. The machine that dispenses salad dressings is working properly when 8 ounces are dispensed. Suppose that the average amount dispensed in a particular sample of 35 bottles is 7.91 ounces with a variance of 0.03 ounces squared, s^2. Is there evidence that the machine should be stopped and production wait for repairs? The lost production from a shutdown is potentially so great that management feels that the level of confidence in the analysis should be 99%.

Again we will follow the steps in our analysis of this problem.

Answer

STEP 1: Set the null and alternative hypothesis.

The random variable is the quantity of fluid placed in the bottles. This is a continuous random variable and the parameter we are interested in is the mean. Our hypothesis therefore is about the mean. In this case we are concerned that the machine is not filling properly. From what we are told it does not matter if the machine is over-filling or under-filling, both seem to be an equally bad error. This tells us that this is a two-tailed test: if the machine is malfunctioning it will be shutdown regardless if it is from over-filling or under-filling. The null and alternative hypotheses are thus:

$$H_0 : \mu = 8$$

$$Ha : \mu \neq 8$$

STEP 2: Decide the level of significance and draw the graph showing the critical value.

This problem has already set the level of confidence at 99%. The decision seems an appropriate one and shows the thought process when setting the significance level. Management wants to be very certain, as certain as probability will allow, that they are not shutting down a machine that is not in need of repair. To draw the distribution and the critical value, we need to know which distribution to use. Because the sample size is under 100, the appropriate distribution is the *t*-distribution and the relevant critical value is 2.750 from the *t*-table at 0.005 column and 30 degrees of freedom (closest available row to our actual 34 *df* here). We need to draw the graph and mark these points.

STEP 3: Calculate sample parameters and the test statistic.

The sample parameters are provided, the sample mean is 7.91 and the sample variance is .03 and the sample size is 35. We need to note that the sample variance was provided, not the sample standard deviation, which is what we need for the formula. Remembering that the standard deviation is simply the square root of the variance, we therefore know the sample standard deviation, s, is 0.173. With this information we can calculate the test statistic as -3.07, and mark it on the graph.

$$t_{obs} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{7.91 - 8}{\cdot 173/\sqrt{35}} = -3.07$$

STEP 4: Compare test statistic and the critical values.

Now we compare the test statistic and the critical value by placing the test statistic on the graph. The test statistic is in the tail, decidedly greater than the critical value of 2.750. We note that even the very small difference between the hypothesized value and the sample value is still a large number of standard errors. The sample mean is only 0.08 ounces different from the required level of 8 ounces, but it is 3+ standard errors away from the required 8 ounces, and thus we reject the null hypothesis.

STEP 5: Reach a conclusion.

Three standard errors of a test statistic will guarantee that the test will fail. The probability that anything is beyond three standard errors of a hypothesized null value - given a large enough sample size - is close to zero. Looking at the closest *t*-scores in *df*=30 row in the *t*-table, we get the p-value interval of (.01, .002) after doubling the one-tailed probabilities of .005 and .001. Our formal conclusion would be “At a 99% level of confidence, we reject the null hypothesis that the sample mean came from a distribution with a mean of 8 ounces”. Or less formally, and getting to the point, “At a 99% level of confidence, we conclude that the machine is under-filling the bottles and is in need of repair”.

Hypothesis Test for Proportions

Just as there were confidence intervals for proportions, or more formally, the population parameter P, there is the ability to test hypotheses concerning P.

The estimated value (point estimate) for P is P' where $P' = x/n$, x is the number of observations in the category of interest in the sample and n is the sample size.

When you perform a hypothesis test of a population proportion P, you take a random sample from the population. To ensure normality of the distribution, sampling must be random and the total sample size must be greater than 100. There is no distribution that can correct for this small sample bias and thus if these conditions are not met we simply cannot test the hypothesis with the data available at that time. We met this condition when we were first estimating confidence intervals for P.

Again, we begin with the modified standardizing formula:

$$z = \frac{P' - P}{\sqrt{\frac{P(1-P)}{n}}}$$

Substituting P_0, the hypothesized value of P, we have:

$$z_{obs} = \frac{P' - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

This is the test statistic for testing hypothesized values of P, where the null and alternative hypotheses take one of the following forms:

Two-tailed test	One-tailed test	One-tailed test
$H_0 : P = P_0$	$H_0 : P \leq P_0$	$H_0 : P \geq P_0$
$H_a : P \neq P_0$	$H_a : P > P_0$	$H_a : P < P_0$

Table 7.5.1

The decision rule stated above applies here also: if the calculated value of z_{obs} shows that the sample proportion is "too many" standard errors from the hypothesized proportion, the null hypothesis is rejected. The decision as to what is "too many" is pre-determined by the analyst depending on the level of significance required in the test.

Example 7.5.4

The mortgage department of a large bank is interested in the nature of loans of first-time borrowers. This information will be used to tailor their marketing strategy. They believe that 50% of first-time borrowers take out smaller loans than other borrowers. They perform a hypothesis test to determine if the percentage is **different from 50%**. They sample **101 first-time borrowers** and find **54** of these loans are smaller that the other borrowers. For the hypothesis test, they choose a 5% level of significance.

Answer

STEP 1: Set the null and alternative hypothesis.

$H_0 : P = 0.50 \qquad H_a : P \neq 0.50$

The words **"is different from"** tell you this is a two-tailed test. The Type I and Type II errors are as follows: The Type I error is to conclude that the proportion of borrowers is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true). The Type II error is there is not enough evidence to conclude that the proportion of first time borrowers differs from 50% when, in fact, the proportion does differ from 50%. (You fail to reject the null hypothesis when the null hypothesis is false.)

STEP 2: Decide the level of significance and draw the graph showing the critical value

The level of confidence has been set by the problem at 95%. Because this is two-tailed test one-half of the α value will be in the upper tail and one-half in the lower tail as shown on the graph. The critical value for the normal distribution at the 95% level of confidence is 1.96. This can easily be found on the Student's *t*-table at the very bottom at infinite degrees of freedom remembering that at infinity the *t*-distribution is the normal distribution. Of course, the value can also be found on the standard normal table but you have go looking for the tail probability, $\alpha/2$, inside the body of the table and then read out to the sides and top for the number of standard errors.

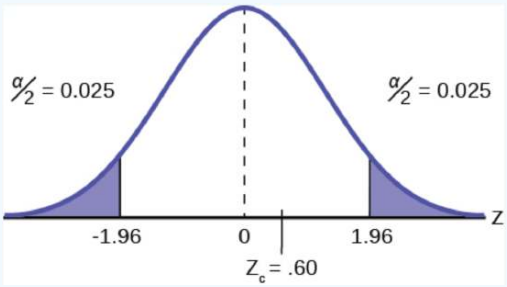


Figure 7.5.3

STEP 3: Calculate the sample parameters and critical value of the test statistic.

The test statistic is a normal distribution, z, for testing proportions and is:

$$z = \frac{P' - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{.53 - .50}{\sqrt{\frac{.5(.5)}{101}}} = 0.60$$

For this case, the sample of 101 found 54 first-time borrowers were different from other borrowers. The sample proportion, $P' = 54/101 = 0.53$ The test question, therefore, is : "Is 0.53 significantly different from 0.50?" Putting these values into the formula for the test statistic we find that 0.53 is only 0.60 standard errors away from 0.50. This is barely off of the mean of the standard normal distribution of zero. There is virtually no difference from the sample proportion and the hypothesized proportion in terms of standard errors.

STEP 4: Compare the test statistic and the critical value.

The observed value is well within the critical values of ± 1.96 standard errors and thus we cannot reject the null hypothesis. To reject the null hypothesis we need significant evidence of difference between the hypothesized value and the sample value. In this case the sample value is very nearly the same as the hypothesized value measured in terms of standard errors.

STEP 5: Reach a conclusion.

The formal conclusion would be "At a 95% level of confidence we cannot reject the null hypothesis that 50% of first-time borrowers have the same size loans as other borrowers". Less formally, we would say that "There is no evidence that one-half of first-time borrowers are significantly different in loan size from other borrowers". Notice the length to which the conclusion goes to include all of the conditions that are attached to the conclusion. Statisticians, for all the criticism they receive, are careful to be very specific even when this seems trivial. Statisticians cannot say more than they know and the data constrain the conclusion to be within the metes and bounds of the data.

Exercise 7.5.3

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 104 students and 89 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

Example 7.5.5

Suppose a consumer group suspects that the proportion of households that have three or more cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test using 90% confidence. Their marketing people survey 150 households with the result that 43 of the households have three or more cell phones.

Answer

Here is an abbreviated version of the system to solve hypothesis tests applied to a test on a proportions.

$$H_0 : P = 0.3$$

$$H_a : P \neq 0.3$$

$$n = 150$$

$$P' = \frac{x}{n} = \frac{43}{150} = 0.287$$

$$z_{obs} = \frac{P' - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0.287 - 0.3}{\sqrt{\frac{.3(.7)}{150}}} = 0.347$$

At a confidence level of 90% we cannot reject the null hypothesis that the consumer group is correct.

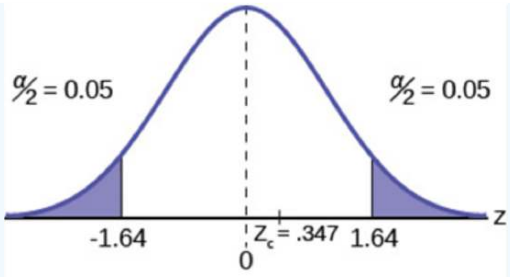


Figure 7.5.4

Example 7.5.6

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Answer

We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be:

$H_0 : P \leq 0.00034$ $H_a : P > 0.00034$

If we commit a Type I error, we are essentially accepting an incorrect claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.

This page titled 7.5: Full Hypothesis Test Examples is shared under a CC BY license and was authored, remixed, and/or curated by .

7.6: Chapter 7 Key Terms

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$. If the size n of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distribution regardless of the shape of the population. The expected value of the mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Critical Value

The t or z value set by the researcher that measures the probability of a Type I error, α.

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Normal Distribution

For a continuous random variable (RV), where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean, on average; notation: s for sample standard deviation and σ for population standard deviation.

Standard Error of the Mean

The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Student's *t*-Distribution

Investigated and reported by William Sealy Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- The distribution is symmetrical about its mean of zero. However, it is more spread out and flatter than the normal distribution.
- It approaches the standard normal distribution as n gets larger. When $n \geq 100$, the *t*-distribution and standard normal distribution are interchangeable.
- There is a "family" of *t*-distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of observations.

Test Statistic

The formula that counts the number of standard errors on the relevant distribution that estimated parameter is away from the hypothesized null value.

Type I Error

Type I Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type II Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

This page titled 7.6: Chapter 7 Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by .

7.7: Chapter 7 Review

7.1 Null and Alternative Hypotheses

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality (=, ≤ or ≥)
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using not equal, less than or greater than symbols, i.e., (≠, <, or >).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. **Never** state that a claim is **proven** true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

7.2 Outcomes and the Type I and Type II Errors

A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1–\beta$, quantifies the likelihood that a test will yield the correct result of a false null hypothesis being rejected. A high power is desirable.

7.3 Distribution Needed for Hypothesis Testing

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a random sample and the population is approximately normally distributed, or the sample size is large.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a random sample with $n > 100$.

7.4 Full Hypothesis Test Examples

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph and calculate the test statistic.
5. Compare the calculated test statistic with the z critical value determined by the level of significance required by the test and make a decision (cannot reject H_0 or reject H_0), and write a clear conclusion using full sentences.

7.7: Chapter 7 Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.8: Chapter 7 Formula Review

7.3 Distribution Needed for Hypothesis Testing

Sample size	Test statistic
< 100 (σ unknown)	$t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
< 100 (σ known)	$z_{obs} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
> 100 (σ unknown)	$z_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
> 100 (σ known)	$z_{obs} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

Table 7.8.1. Test Statistics for Test of Means, Varying Sample Size, Population Known or Unknown

7.8: Chapter 7 Formula Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.9: Chapter 7 Homework

7.2 Null and Alternative Hypotheses

1. 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.

2. 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.

3. The American family has an average of two children. What is the random variable? Describe in words.

4. 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.

5. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting 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.

6. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

7. In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

8. 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.

1. H_0: _____
2. H_a: _____

9. 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?

1. H_0: _____
2. H_a: _____

10. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven 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?

1. H_0: _____
2. H_a: _____

7.3 Outcomes and Type I and Type II Errors

11. 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.

12. 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.

13. For Exercise 7.9.12, what are α and β in words?

14. In words, describe $1–\beta$ for Exercise 7.9.12.

15. 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.

16. 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. Which is the error with the greater consequence?

17. The power of a test is 0.981. What is the probability of a Type II error?

18. 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.

19. 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, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

20. 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?

7.4 Distribution Needed for Hypothesis Testing

21. Which two distributions can you use for hypothesis testing for this chapter?

22. Which distribution do you use when you are testing a population mean and the population standard deviation is known? Assume sample size is large. Assume a normal distribution with $n \geq 100$.

23. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume a normal distribution, with $n \geq 100$.

24. 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.

25. 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?

26. It is thought that 42% of respondents in a taste test would prefer Brand A. In a particular test of 100 people, 39% preferred Brand A. What distribution should you use to perform a hypothesis test?

27. 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?

28. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. The data are not from a random sample. Can you accurately perform the hypothesis test?

7.5 Full Hypothesis Test Examples

29. Assume $H_0 : \mu \geq 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

30. Assume $H_0 : \mu \leq 6$ and $H_a : \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

31. Assume $H_0 : P = 0.25$ and $H_a : P \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

32. Draw the general graph of a left-tailed test.

33. Draw the graph of a two-tailed test.

34. A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

35. Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

36. A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

37. You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

38. If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

39. Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

40. Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

41. Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

7.2 Null and Alternative Hypotheses

42. 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).

1. The mean number of years Americans work before retiring is 34.
2. At most 60% of Americans vote in presidential elections.
3. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
4. Twenty-nine percent of high school seniors get drunk each month.
5. Fewer than 5% of adults ride the bus to work in Los Angeles.
6. The mean number of cars a person owns in her lifetime is not more than ten.
7. About half of Americans prefer to live away from cities, given the choice.
8. Europeans have a mean paid vacation each year of six weeks.
9. The chance of developing breast cancer is under 11% for women.
10. Private universities' mean tuition cost is more than $20,000 per year.

43. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) 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. $P \leq 0.20$

44. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is 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 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. $H_0 : \mu \leq 4.5, H_a : \mu > 4.5$

7.3 Outcomes and Type I and Type II Errors

45. State the Type I and Type II errors in complete sentences given the following statements.

1. The mean number of years Americans work before retiring is 34.
2. At most 60% of Americans vote in presidential elections.
3. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
4. Twenty-nine percent of high school seniors get drunk each month.
5. Fewer than 5% of adults ride the bus to work in Los Angeles.
6. The mean number of cars a person owns in his or her lifetime is not more than ten.
7. About half of Americans prefer to live away from cities, given the choice.

8. Europeans have a mean paid vacation each year of six weeks.
9. The chance of developing breast cancer is under 11% for women.
10. Private universities mean tuition cost is more than $20,000 per year.

46. For statements 1-10 in Exercise 7.9.45, answer the following in complete sentences.

1. State a consequence of committing a Type I error.
2. State a consequence of committing a Type II error.

47. 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 (FDA) 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.

48. 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 184 of her students and finds that 241 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%.

49. 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 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

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

50. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is 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. When conducting 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 mean hours per week currently is 4.5, when in fact, it is higher
d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

7.4 Distribution Needed for Hypothesis Testing

51. It is believed that Lake Wobegon College (LWC) Statistics students get less than seven hours of sleep per night, on average. A survey of 22 LWC Statistics students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, a researcher wants to test whether LWC Statistics students get less than seven hours of sleep per night, on average. Which of the following is the appropriate distribution to be used for this test?

a. $N\left(7.24, \frac{1.93}{\sqrt{22}}\right)$

b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

7.5 Full Hypothesis Test Examples

52. A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using alpha = 0.05, is the data highly inconsistent with the claim?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

53. From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

54. The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve newspapers yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

55. An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

56. The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let X = the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten (at the 5% level)?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

57. In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased.

81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

58. Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own (at a 5% level). You randomly survey 164 of her past Elementary Statistics students and find that 84 feel more enriched as a result of her class. Now, what do you think?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

59. A Nissan Motor Corporation advertisement read, “The average man’s I.Q. is 107. The average brown trout’s I.Q. is 4. So why can’t man catch brown trout?” Suppose you believe that the brown trout’s mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief and use 95% confidence level.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

60. Refer to Exercise 7.9.59. Conduct a hypothesis test at 95% confidence level to see if your decision and conclusion would change if your belief were that the brown trout’s mean I.Q. is *not* four.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

61. According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don’t believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7 at 1% level?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

62. A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 126 Americans surveyed, only four had seen or sensed the presence of an angel. As a result of the contingent’s survey, would you agree with the *Newsweek* poll at 5% level? In complete sentences, also give three reasons why the two polls might give different results.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

63. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours (at 10% level)?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

64. Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California's percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California was randomly selected for comparison. In the same year, 80 of the 104 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

65. According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Eleven out of 120 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test at 5% level to determine if the rate is still 14% or if it has decreased.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

66. The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test at 5% level.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

67. Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test at 5% level.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

68. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.
At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.

b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

69. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is 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. At a significance level of $\alpha = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75
b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

70. According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

71. A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

72. Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Out of 145 randomly selected fatal accidents are examined, it is determined that 75 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

73. The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

74. For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The

class randomly selected 104 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

75. The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

76. A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

77. A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

3; 2; 1; 3; 7; 2; 9; 4; 6; 6; 8; 0; 5; 6; 4; 2; 1; 3; 4; 1

At the $\alpha = 0.05$ level can it be concluded that the sample mean is higher than 5.8 visits per year?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

78. According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:
5; 4; 5; 4; 4; 3; 6; 4; 3; 3; 5; 5; 6; 3; 3; 2; 7; 4; 5; 2; 2; 2; 3; 2
At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using p-values. Interpret the p-value.

79. The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.

d. Test the above hypotheses using p-values. Interpret the p-value.

7.9: Chapter 7 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.10: Chapter 7 Solutions

1. The random variable is the mean Internet speed in Megabits per second.

3. The random variable is the mean number of children an American family has.

5. The random variable is the proportion of people picked at random in Times Square visiting the city.

7.

1. $H_0 : P \geq 0.42$
2. $H_a : P < 0.42$

9.

1. $H_0 : \mu = 15$
2. $H_a : \mu \neq 15$

11.

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

13.

α = the probability that you think the bag cannot withstand -15 degrees F, when in fact it can

β = the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

15.

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

17. 0.019

19. 0.998

21. A normal distribution or a Student's t-distribution

23. Use a Student's t-distribution

25. a normal distribution for a single population mean

27. It must be approximately normally distributed.

29. This is a left-tailed test.

31. This is a two-tailed test.

33.

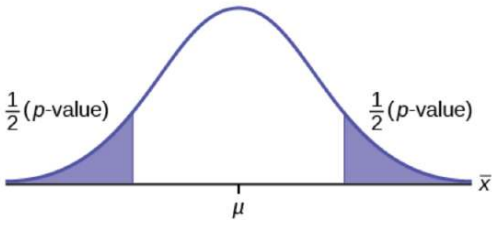


Figure 7.10.1

35. a right-tailed test

37. a left-tailed test

39. This is a left-tailed test.

41. This is a two-tailed test.

42.

1. $H_0 : \mu = 34; H_a : \mu \neq 34$
2. $H_0 : P \leq 0.60; H_a : P > 0.60$
3. $H_0 : \mu \geq 100,000; H_a : \mu < 100,000$
4. $H_0 : P = 0.29; H_a : P \neq 0.29$
5. $H_0 : P \geq 0.05; H_a : P < 0.05$
6. $H_0 : \mu \leq 10; H_a : \mu > 10$
7. $H_0 : P = 0.50; H_a : P \neq 0.50$
8. $H_0 : \mu = 6; H_a : \mu \neq 6$
9. $H_0 : P \geq 0.11; H_a : P < 0.11$
10. $H_0 : \mu \leq 20,000; H_a : \mu > 20,000$

43. c

45.

1. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
2. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
3. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
4. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
5. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer that 5% do.
6. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
7. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
8. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.
9. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
10. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

47. b

49. d

51. d

52.

$H_0 : \mu \geq 50,000$

$H_a : \mu < 50,000$

$\overline{x}$ = the average lifespan of a brand of tires.

$[43,537,49,463]$

Decision: Reject the null hypothesis.

Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.

54.

$H_0 : \mu = \$1.00$

$H_a : \mu \neq \$1.00$

$\overline{x}$ = the average cost of a daily newspaper.

$[\$0.80, \$1.10]$

Decision: Do not reject the null hypothesis.

Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is $1. The mean cost could be $1.

56.

$H_0 : \mu = 10$

$H_a : \mu \neq 10$

$\overline{x}$ = the mean number of sick days an employee takes per year.

$[4.9443, 11.806]$

Decision: Do not reject the null hypothesis.

Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.

60.

$H_0 : \mu = 4$

$H_a : \mu \neq 4$

$\overline{x}$ = the average I.Q. of a set of brown trout.

$[3.8865, 5.9468]$

Decision: Reject the null hypothesis.

Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.

67.

$H_0 : \mu \leq 69,110$

$H_a : \mu > 69,110$

$\overline{x}$ = the mean salary in dollars for California registered nurses.

$[\$68,757, \$73,485]$

Decision: Reject the null hypothesis.

Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.

69. c

71.

$H_0 : P = 0.488$

$H_a : P \neq 0.488$

Decision: Reject the null hypothesis.

Conclusion: At the 5% level of significance, there is not enough evidence to conclude that 48.8% of families own stocks. The survey does not appear to be accurate.

73.

$H_0 : P = 0.517$

$H_a : P \neq 0.517$

Decision: Do not reject the null hypothesis.

Conclusion: At the 5% significance level, there is enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.

However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

75.

$H_0 : \mu \geq 11.52$

$H_a : \mu < 11.52$

Decision: Reject the null hypothesis.

Conclusion: At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.

We would make the same conclusion if alpha was 1% because the p-value is almost 0.

77.

$H_0 : \mu \leq 5.8$

$H_a : \mu > 5.8$

Decision: Do not reject the null hypothesis.

Conclusion: At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

79.

$H_0 : \mu \geq 150$

$H_a : \mu < 150$

Decision: Do not reject the null hypothesis.

Conclusion: At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average. The student academic group's claim appears to be correct.

7.10: Chapter 7 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.11: Chapter 7 References

7.2 Null and Alternative Hypotheses

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

7.5 Full Hypothesis Test Examples

Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.

Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.

Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).

Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).

Data from *Growing by Degrees* by Allen and Seaman.

Data from La Leche League International. Available online at http://www.lalecheleague.org/Law/BAFeb01.html.

Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).

Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).

Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.

Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)

Data from the U.S. Census Bureau, available online at http://quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).

Data from the United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/.

Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.

Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).

Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).

“Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at http://research.fhda.edu/factbook/DA...t_da_2006w.pdf.

Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).

Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at http://www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

7.11: Chapter 7 References is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

8: Hypothesis Testing with Two Samples

8.1: Introduction
8.2: Comparing Two Independent Population Means
8.3: Cohen's Standards for Small, Medium, and Large Effect Sizes
8.4: Comparing Two Independent Population Proportions
8.5: Matched or Paired Samples
8.6: Chapter 8 Key Terms
8.7: Chapter 8 Review
8.8: Chapter 8 Formula Review
8.9: Chapter 8 Homework
8.10: Chapter 8 Solutions
8.11: Chapter 8 References

This page titled 8: Hypothesis Testing with Two Samples is shared under a CC BY license and was authored, remixed, and/or curated by .

8.1: Introduction



Figure 8.1.1 If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies 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. Many business applications require comparing two groups. It may be the investment returns of two different investment strategies, or the differences in production efficiency of different management styles.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or **matched pairs**. **Independent groups** consist of two samples that are independent; that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions of each group.

This page titled 8.1: Introduction is shared under a CC BY license and was authored, remixed, and/or curated by .

8.2: Comparing Two Independent Population Means

The comparison of two independent population means is very common and provides a way to test the hypothesis that the two groups differ from each other. Is the night shift less productive than the day shift, are the rates of return from fixed asset investments different from those from common stock investments, and so on? An observed difference between two sample means depends on both the means and the sample standard deviations. Very different means can occur by chance if there is great variation among the individual samples. The test statistic will have to account for this fact.

When we developed the hypothesis test for the mean and proportions we began with the Central Limit Theorem. We recognized that a sample mean came from a distribution of sample means, and sample proportions came from the sampling distribution of sample proportions. This made our sample parameters, the sample means and sample proportions, into random variables. It was important for us to know the distribution that these random variables came from. The Central Limit Theorem gave us the answer: the normal distribution. Our Z and t statistics came from this theorem. This provided us with the solution to our question of how to measure the probability that a sample mean came from a distribution with a particular hypothesized value of the mean or proportion. In both cases that was the question: what is the probability that the mean (or proportion) from our sample data came from a population distribution with the hypothesized value we are interested in?

Now we are interested in whether or not two samples have the same mean. Our question has not changed: Do these two samples come from the same population distribution? To approach this problem we create a new random variable. We recognize that we have two sample means, one from each set of data, and thus we have two random variables coming from two unknown distributions. To solve the problem we create a new random variable, the difference between the sample means. This new random variable also has a distribution and, again, the Central Limit Theorem tells us that this new distribution is normally distributed, regardless of the underlying distributions of the original data. A graph may help to understand this concept.

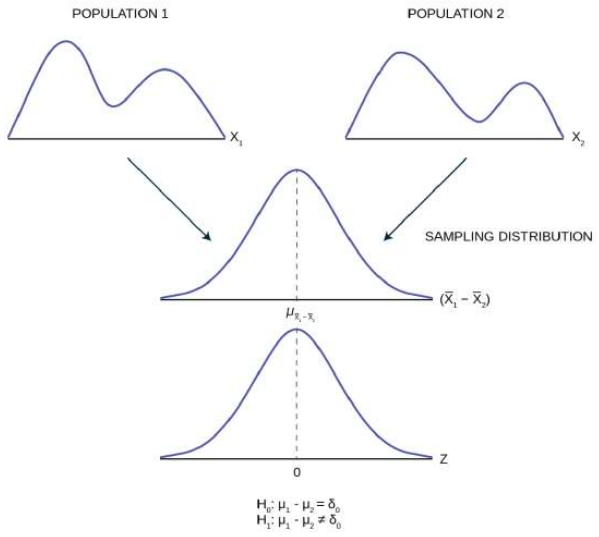


Figure 8.2.1

Pictured are two distributions of data, X_1 and X_2, with unknown means and standard deviations. The second panel shows the sampling distribution of the newly created random variable $(\overline{x}_1 - \overline{x}_2)$. This distribution is the theoretical distribution of many many sample means from population 1 minus sample means from population 2. The Central Limit Theorem tells us that this theoretical sampling distribution of differences in sample means is normally distributed, regardless of the distribution of the actual population data shown in the top panel. Because the sampling distribution is normally distributed, we can develop a standardizing formula and calculate probabilities from the standard normal distribution in the bottom panel, the Z distribution. We have seen this same analysis before in Chapter 5.2.

The Central Limit Theorem, as before, provides us with the standard deviation of the sampling distribution, and further, that the expected value of the mean of the distribution of differences in sample means is equal to the differences in the population means. Mathematically this can be stated:

$$E\left(\mu_{\overline{x}_1} - \mu_{\overline{x}_2}\right) = \mu_1 - \mu_2$$

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of **the difference in sample means**, $\overline{x}_1 - \overline{x}_2$.

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

We remember that substituting the sample variance for the population variance when we did not have the population variance was the technique we used when building the confidence interval and the test statistic for the test of hypothesis for a single mean back in Confidence Intervals and Hypothesis Testing with One Sample. **The test statistic (*t*-score) is calculated as follows:**

$$t_{obs} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively and
- σ_1 and σ_2 are the unknown population standard deviations.
- μ_1 and μ_2 are the unknown population means.
- $\overline{x}_1$ and $\overline{x}_2$ are the sample means.

The number of **degrees of freedom (*df*)** requires a somewhat complicated calculation. The df are not always a whole number. The test statistic above is approximated by the Student's t-distribution with df equal to n_1 + n_2 - 2.

The format of the sampling distribution, differences in sample means, specifies that the format of the null and alternative hypothesis is:

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

$$H_a : \mu_1 - \mu_2 \neq \delta_0$$

where δ_0 is the hypothesized difference between the two means. If the question is simply "is there any difference between the means?" then $\delta_0 = 0$ and the null and alternative hypotheses becomes:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

An example of when δ_0 might not be zero is when the comparison of the two groups requires a specific difference for the decision to be meaningful. Imagine that you are making a capital investment. You are considering changing from your current model machine to another. You measure the productivity of your machines by the speed they produce the product. It may be that a contender to replace the old model is faster in terms of product throughput, but is also more expensive. The second machine may also have more maintenance costs, setup costs, etc. The null hypothesis would be set up so that the new machine would have to be better than the old one by enough to cover these extra costs in terms of speed and cost of production. This form of the null and alternative hypothesis shows how valuable this particular hypothesis test can be. For most of our work we will be testing simple hypotheses asking if there is any difference between the two distribution means.

Example 8.2.1

The Kona Iki Corporation produces coconut milk. They take coconuts and extract the milk inside by drilling a hole and pouring the milk into a vat for processing. They have both a day shift (called the B shift) and a night shift (called the G shift) to do this part of the process. They would like to know if the day shift and the night shift are equally efficient in processing the coconuts. A study is done sampling 9 shifts of the G shift and 16 shifts of the B shift. The results of the number of hours required to process 100 pounds of coconuts is presented in the table below.

	Sample Size	Average Number of Hours to Process 100 Pounds of Coconuts	Sample Standard Deviation
G Shift	9	2	0.866
B Shift	16	3.2	1.00

Table 8.2.1

Is there a difference in the mean amount of time for each shift to process 100 pounds of coconuts? Test at the 5% level of significance.

Answer

The population standard deviations are not known and cannot be assumed to equal each other. Let g be the subscript for the G Shift and b be the subscript for the B Shift. Then, μ_g is the population mean for G Shift and μ_b is the population mean for B Shift. This is a test of two **independent groups**, two population **means**.

Random variable: $\overline{x}_g - \overline{x}_b$ = difference in the sample mean amount of time between the G Shift and the B Shift takes to process the coconuts.

$H_0 : \mu_g = \mu_b$ $\quad$ $H_0 : \mu_g - \mu_b = 0$

$H_a : \mu_g \neq \mu_b$ $\quad$ $H_a : \mu_g - \mu_b \neq 0$

The words **"the same"** tell you H_0 has an "=". Since there are no other words to indicate H_a, is either faster or slower. This is a two tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula above: n_1 + n_2 - 2 = 9 + 16 - 2 = 23.

We next find the critical value on the t-table using the degrees of freedom from above. The critical value, 2.069, is found in the .025 column, this is $\alpha/2$, at 23 degrees of freedom. Next we calculate the test statistic.

$$t_{obs} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -3.01$$

Make a decision: Since the calculated t-value is in the tail we reject the null hypothesis that there is no difference between the two groups. The means are different.

The calculated difference in the two means is -1.2 and this is 3.01 standard errors from the mean.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that the G Shift takes to process 100 pounds of coconuts is different from the B Shift (mean number of hours for the B Shift is greater than the mean number of hours for the G Shift).

Note

When the sum of the sample sizes is larger than $100\,(n_1 + n_2 > 100)$ you can use the normal distribution to approximate the Student's t.

Example 8.2.2

A study is done to determine if Company A retains its workers longer than Company B. It is believed that Company A has a higher retention than Company B. The study finds that in a sample of 11 workers at Company A their average time with the company is four years with a standard deviation of 1.5 years. A sample of 9 workers at Company B finds that the average time with the company was 3.5 years with a standard deviation of 1 year. Test this proposition at the 1% level of significance.

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?

f. Is this test right-, left-, or two-tailed?

g. What is the value of the test statistic?

h. Can you reject the null hypothesis?

i. Conclusion:

Answer

a. two means because time is a continuous random variable

b. unknown

c. Student's t

d. $\overline{x}_A - \overline{x}_B$

e. $H_0 : \mu_A \leq \mu_B$; $H_a : \mu_A > \mu_B$

f. right one-tailed test

g.

$t_{obs} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 0.89$

h. Cannot reject the null hypothesis that there is no difference between the two groups. Test statistic is not in the tail. The critical value of the *t* distribution is 2.552 with 18 degrees of freedom. This example shows how difficult it is to reject a null hypothesis with a very small sample. The critical values require very large test statistics to reach the tail.

i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the retention of workers at Company A is longer than Company B, on average.

Example 8.2.3

An interesting research question is the effect, if any, that different types of teaching formats have on the grade outcomes of students. To investigate this issue one sample of students' grades was taken from a hybrid class and another sample taken from a standard lecture format class. Both classes were for the same subject. The mean course grade in percent for the 35 hybrid students is 74 with a standard deviation of 16. The mean grades of the 40 students form the standard lecture class was 76 percent with a standard deviation of 9. Test at 5% to see if there is any significant difference in the population mean grades between standard lecture course and hybrid class.

Answer

We begin by noting that we have two groups, students from a hybrid class and students from a standard lecture format class. We also note that the random variable, what we are interested in, is students' grades, a continuous random variable. We could have asked the research question in a different way and had a binary random variable. For example, we could have studied the percentage of students with a failing grade, or with an A grade. Both of these would be binary and thus a test of proportions and not a test of means as is the case here. Finally, there is no presumption as to which format might lead to higher grades so the hypothesis is stated as a two-tailed test.

$H_0 : \mu_1 = \mu_2$
$H_a : \mu_1 \neq \mu_2$

As would virtually always be the case, we do not know the population variances of the two distributions and thus our test statistic is:

$$t_{obs} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{(74 - 76) - 0}{\sqrt{\frac{16^2}{35} + \frac{9^2}{40}}} = -0.65$$

To determine the critical value of the Student's *t* we need the degrees of freedom. For this case we use: $df = n_1 + n_2 - 2 = 35 + 40 - 2 = 73$. Using the closest available row in our *t* table, for 80 df, we can determine that $t_{\alpha/2} = 1.990$. Again as always we determine if the calculated value is in the tail determined by the critical value. In this case we do not even need to look up the critical value: the calculated value of the difference in these two average grades is not even one standard deviation apart. Certainly not in the tail.

Conclusion: Cannot reject the null at $\alpha = 5\%$. Therefore, there is not sufficient evidence that the grades in hybrid and standard classes differ.

This page titled 8.2: Comparing Two Independent Population Means is shared under a CC BY license and was authored, remixed, and/or curated by .

8.3: Cohen's Standards for Small, Medium, and Large Effect Sizes

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	0.2 - 0.8
Large	≥ 0.8

Table 8.3.1 Cohen's Standard Effect Sizes

Cohen's d is the measure of the standardized difference between two means, calculated as follows:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{2} + \frac{s_2^2}{2}}}$$

It is important to note that Cohen's d does not provide a level of confidence as to the magnitude of the size of the effect comparable to the other tests of hypothesis we have studied. The sizes of the effects are simply indicative.

Example 8.3.1

Calculate Cohen's d for the difference between two groups, where $\bar{x}_1 = 4$, $s_1 = 1.5$, $n_1 = 11$, and $\bar{x}_2 = 3.5$, $s_2 = 1$, $n_2 = 9$. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Answer

$d = 0.89$

The effect is large because 0.89 is greater than Cohen's value of 0.8 for large effect sizes. The size of the differences of the means for the two companies is large, indicating that there is a meaningful difference between them.

This page titled 8.3: Cohen's Standards for Small, Medium, and Large Effect Sizes is shared under a CC BY license and was authored, remixed, and/or curated by .

8.4: Comparing Two Independent Population Proportions

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance in the sampling. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the two population proportions.

Like the case of differences in sample means, we construct a sampling distribution for differences in sample proportions: $(P'_A - P'_B)$ where $P'_A = \frac{x_A}{n_A}$ and $P'_B = \frac{x_B}{n_B}$ are the sample proportions for the two sets of data in question. X_A and X_B are the number of observations in each sample group of interest, respectively, and n_A and n_B are the respective sample sizes from the two groups. Again we go the Central Limit theorem to find the distribution of this sampling distribution for the differences in sample proportions. And again we find that this sampling distribution, like the ones past, are normally distributed as proved by the Central Limit Theorem.

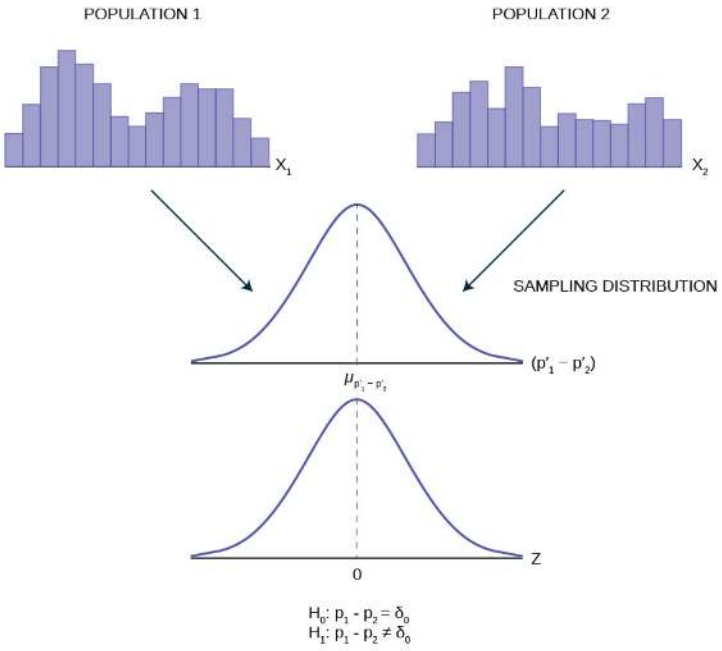


Figure 8.4.1

Generally, the null hypothesis allows for the test of a difference of a particular value, δ_0, just as we did for the case of differences in means.

$$H_0 : P_A - P_B = \delta_0$$

$$H_1 : P_A - P_B \neq \delta_0$$

The confidence interval formula is:

$$(P'_2 - P'_1) \pm z_{\frac{\alpha}{2}} * \sqrt{\frac{P'_1 * (1 - P'_1)}{n_1} + \frac{P'_2 * (1 - P'_2)}{n_2}}$$

Example 8.4.1

A bank has recently acquired a new branch and thus has customers in this new territory. They are interested in the default rate in their new territory. They wish to test the hypothesis that the default rate is different from their current customer base. They sample 200 files in area A, their current customers, and find that 20 have defaulted. In area B, the new customers, another sample of 200 files shows 12 have defaulted on their loans. At a 1% level of significance can we say that the default rates are the same or different?

Answer

This is a test of proportions. We know this because the underlying random variable is binary, default or not default. Further, we know it is a test of differences in proportions because we have two sample groups, the current customer base and the newly acquired customer base. Let A and B be the subscripts for the two customer groups. Then $p_A$and p_B are the two population proportions we wish to test.

Random Variable:

$P'_A - P'_B$ = difference in the proportions of customers who defaulted in the two groups.

$H_0 : P_A = P_B$

$H_a : P_A \neq P_B$

The words "is different" tell you the test is two-tailed.

Distribution for the test:

Estimated proportion for group A: $P'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $P'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

The estimated confidence interval for the difference between the two groups is:

$$(P'_2 - P'_1) \pm z_{\frac{\alpha}{2}} * \sqrt{\frac{P'_1 * (1 - P'_1)}{n_1} + \frac{P'_2 * (1 - P'_2)}{n_2}} = [-.03, .11]$$

Make a decision: Since the calculated confidence interval contains 0, we cannot 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 between the proportions of customers who defaulted in the two groups.

Exercise 8.4.1

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Six out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.

This page titled 8.4: Comparing Two Independent Population Proportions is shared under a CC BY license and was authored, remixed, and/or curated by .

8.5: Matched or Paired Samples

In most cases of economic or business data we have little or no control over the process of how the data are gathered. In this sense the data are not the result of a planned controlled experiment. In some cases, however, we can develop data that are part of a controlled experiment. This situation occurs frequently in quality control situations. Imagine that the production rates of two machines built to the same design, but at different manufacturing plants, are being tested for differences in some production metric such as speed of output or meeting some production specification such as strength of the product. The test is the same in format to what we have been testing, but here we can have matched pairs for which we can test if differences exist. Each observation has its matched pair against which differences are calculated. First, the differences in the metric to be tested between the two lists of observations must be calculated, and this is typically labeled with the letter "d." Then, the average of these matched differences, $\overline{x}_d$ is calculated as is its standard deviation, s_d. We expect that the standard deviation of the differences of the matched pairs will be smaller than unmatched pairs because presumably fewer differences should exist because of the correlation between the two groups.

When using a hypothesis test for matched or paired samples, the following characteristics may be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student's *t*-test for a single population mean with n–1 degrees of freedom, where n is the number of differences, that is, the number of pairs not the number of observations.

The null and alternative hypotheses for this test are:

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

The test statistic is:

$$t_{obs} = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example 8.5.1

A company has developed a training program for its entering employees because they have become concerned with the results of the six-month employee review. They hope that the training program can result in better six-month reviews. Each trainee constitutes a “pair”, the entering score the employee received when first entering the firm and the score given at the six-month review. The difference in the two scores were calculated for each employee and the means for before and after the training program was calculated. The sample mean before the training program was 20.4 and the sample mean after the training program was 23.9. The standard deviation of the differences in the two scores across the 20 employees was 3.8 points. Test at the 5% significance level the null hypothesis that the training program makes no difference or worsens employees’ scores against the alternative that the training program helps improve the employees’ scores.

Answer

The first step is to identify this as a two sample case: before the training and after the training. This differentiates this problem from simple one sample issues. Second, we determine that the two samples are "paired." Each observation in the first sample has a paired observation in the second sample. This information tells us that the null and alternative hypotheses should be:

$$H_0 : \mu_d \leq 0$$

$$H_a : \mu_d > 0$$

This form reflects the implied claim that the training course improves scores; the test is one-tailed and the claim is in the alternative hypothesis. Because the experiment was conducted as a matched paired sample rather than simply taking scores from people who took the training course those who didn't, we use the matched pair test statistic:

$$\text{Test Statistic: } t_{obs} = \frac{\overline{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{(23.9 - 20.4) - 0}{\left(\frac{3.8}{\sqrt{20}}\right)} = 4.12$$

In order to solve this equation, the individual scores (not provided here), pre-training course and post-training course would be needed to be used to calculate the standard deviation across the individual differences. From these differences we would calculate the standard deviation across the individual differences as follows:

$$s_d = \sqrt{\frac{\Sigma \left(x_d - \overline{x}_d\right)^2}{n_d - 1}}$$

We can now compare the calculated value of the test statistic, 4.12, with the critical value. The critical value is a Student's *t* with degrees of freedom equal to the number of pairs, not observations, minus 1. In this case 20 pairs and at 95% confidence level $t_{a/2} = 1.729$ at $df = 20 - 1 = 19$. The calculated test statistic is most certainly in the tail of the distribution and thus we reject the null hypothesis that there is no difference from the training program. Evidence seems indicate that the training aids employees in gaining higher scores.

Example 8.5.2

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table below. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. 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

Table 8.5.1

Answer

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

Table 8.5.2

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: $\bar{x}_d = -3.13$ 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: $\bar{x}_d$ = 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 test statistic and look up the critical value using the Student's *t*-distribution: The calculated value of the test statistic is -3.06 and the critical value of the t distribution with 7 degrees of freedom at the 5% level of confidence is -1.895 with a one-tailed (in this case, left-tailed) test.

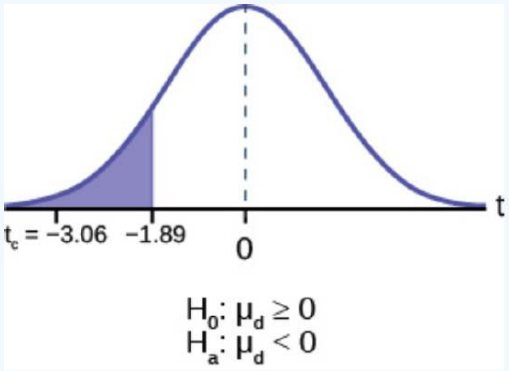


Figure 8.5.1

Compare the critical value for alpha against the calculated test statistic.

The conclusion from using the comparison of the calculated test statistic and the critical value will gives us the result. In this question the calculated test statistic is 3.06 and the critical value is 1.895. The test statistic is clearly in the tail and thus we must reject the null hypotheses that there is no difference between the two situations, hypnotized and not hypnotized.

Make a decision: Reject the null hypothesis, H_0. This means that $\mu_d < 0$ and there is a statistically significant 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.

Example 8.5.3

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

Table 8.5.3

Answer

The coach wants to know if the strength development class makes his players stronger, on average.
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: $\{90, 11, -8, -8\}$.

$\overline{x}_d = 21.3, s_d = 46.7$

Using the difference data, this becomes a test of a single mean.

Define the random variable: $\overline{x}_d$ mean difference in the maximum lift per player.

The distribution for the hypothesis test is a Student's *t* with 3 degrees of freedom.

$H_0: \mu_d \leq 0, H_a: \mu_d > 0$

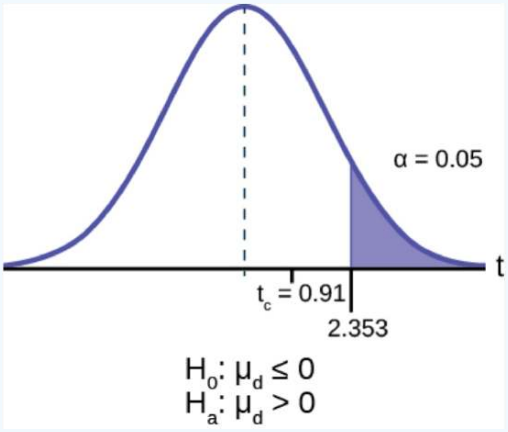


Figure 8.5.2

Calculate the test statistic look up the critical value: The calculated value of the test statistic is 0.91. The critical value of the Student's *t* at 5% level of significance, a one-tailed test, and 3 degrees of freedom is 2.353.

Decision: If the level of significance is 5%, we cannot reject the null hypothesis, because the calculated value of the test statistic is not in the tail.

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.

8.5: Matched or Paired Samples is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.6: Chapter 8 Key Terms

Cohen's *d*

a measure of effect size based on the differences between two means. If d is between 0 and 0.2 then the effect is small. If d is between 0.2 and 0.8, then the effect is medium, and if d exceeds 0.8, then it is a large effect.

Independent Groups

two samples that are selected from two populations, and the values from one population are not related in any way to the values from the other population.

Matched Pairs

two samples that are dependent. Differences between a before and after scenario are tested by testing one population mean of differences.

8.6: Chapter 8 Key Terms is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.7: Chapter 8 Review

8.2 Comparing Two Independent Population Means

Two population means from independent samples where the population standard deviations are not known

- Random Variable: $\overline{x}_1 - \overline{x}_2$ = the difference of the sampling means
- Distribution: Student's *t*-distribution with degrees of freedom = $n_1 + n_2 - 2$

8.3 Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's *d* is a measure of “effect size” based on the differences between two means.

It is important to note that Cohen's d does not provide a level of confidence as to the magnitude of the size of the effect comparable to the other tests of hypothesis we have studied. The sizes of the effects are simply indicative.

8.4 Comparing Two Independent Population Proportions

Test of two population proportions from independent samples.

- Random variable: $P'_A - P'_B$ = difference between the two estimated proportions
- Distribution: normal distribution

8.5 Matched or Paired Samples

A hypothesis test for matched or paired samples (*t*-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random Variable: $\overline{x}_d$ = mean of the differences
- Distribution: Student’s *t*-distribution with n–1 degrees of freedom
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two samples are drawn from the same set of objects.
- Samples are dependent.

8.7: Chapter 8 Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.8: Chapter 8 Formula Review

8.2 Comparing Two Independent Population Means

Standard error: $se = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

Test statistic (*t*-score): $t_{obs} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$

Degrees of freedom: $df = n_1 + n_2 - 2$

8.3 Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's d is the measure of effect size:

$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{2} + \frac{s_2^2}{2}}}$

8.4 Comparing Two Independent Population Proportions

Confidence Interval: $(P_2' - P_1') \pm z_{\frac{\alpha}{2}} * \sqrt{\frac{P_1'*(1-P_1')}{n_1} + \frac{P_2'*(1-P_2')}{n_2}}$

8.5 Matched or Paired Samples

Test Statistic (*t*-score): $t_{obs} = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$

where:

$\overline{x}_d$ is the mean of the sample differences, μ_d is the mean of the population differences, s_d is the sample standard deviation of the differences, and n is the sample size.

8.8: Chapter 8 Formula Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.9: Chapter 8 Homework

8.2 Comparing Two Independent Population Means

Use the following information to answer the next 15 exercises: Indicate if the hypothesis test is for

a. independent group means
b. matched or paired samples
c. single mean
d. two proportions
e. single proportion

1. 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.

2. 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.

3. 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.

4. 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.

5. The average worker in Germany gets eight weeks of paid vacation.

6. According to a television commercial, 80% of dentists agree that Ultrafresh toothpaste is the best on the market.

7. 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.

8. 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?

9. 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?

10. 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.

11. 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.

12. Varsity athletes practice five times a week, on average.

13. 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?

14. 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?

15. 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 non-athletes?

Use the following information to answer the next three exercises: 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.

16. Are standard deviations known or unknown?

17. What is the random variable?

18. Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises: The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, 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 whites and nonwhites.

19. Is this a test of means or proportions?

20. State the null and alternative hypotheses.

1. H_0: _____
2. H_a: _____

21. Is this a right-tailed, left-tailed, or two-tailed test?

22. In symbols, what is the random variable of interest for this test?

23. In words, define the random variable of interest for this test.

24. Which distribution (normal or Student's t) would you use for this hypothesis test?

25. Explain why you chose the distribution you did for the previous question.

26. Calculate the test statistic.

27. 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.

28. At $\alpha = 0.05$, what is your:

1. Decision:
2. Reason for the decision:
3. Conclusion (write out in a complete sentence):

29. Does it appear that the means are the same? Why or why not?

8.4 Comparing Two Independent Population Proportions

Use the following information for the next five exercises. 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.

30. Is this a test of means or proportions?

31. What is the random variable?

32. State the null and alternative hypotheses.

33. Conduct a hypothesis test using the confidence interval approach and 95% confidence. What can you conclude about the two operating systems?

Use the following information to answer the next twelve exercises. In the recent Census, three percent of the U.S. 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. 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.

34. Is this a test of means or proportions?

35. State the null and alternative hypotheses.

1. H_0: _____
2. H_a: _____

36. Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

37. What is the random variable of interest for this test?

38. In words, define the random variable for this test.

39. Which distribution (normal or Student's t) would you use for this hypothesis test?

40. Explain why you chose the distribution you did for the previous question.

41. Calculate the confidence interval. Use 95% confidence.

42. At $\alpha = 0.05$, what is your:

1. Decision:
2. Reason for the decision:
3. Conclusion (write out in a complete sentence):

43. 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?

8.5 Matched or Paired Samples

Use the following information to answer the next five exercises. 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 in the table 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	1

Table 8.9.1

44. What is the random variable?

45. State the null and alternative hypotheses.

46. What conclusion can you draw about the software patch?

a. Test the above hypotheses using confidence intervals. Interpret the CI.
b. Test the above hypotheses using test statistics.

c. Test the above hypotheses using *p*-values. Interpret the *p*-value.

Use the following information to answer next five exercises. 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

Table 8.9.2

47. State the null and alternative hypotheses.

48. What is the sample mean difference?

49. What conclusion can you draw about the juggling class?

a. Test the above hypotheses using confidence intervals. Interpret the CI.
b. Test the above hypotheses using test statistics.
c. Test the above hypotheses using *p*-values. Interpret the *p*-value.

Use the following information to answer the next five exercises. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1% significance level.

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	160	159	166	160	167	169

Table 8.9.3

50. State the null and alternative hypotheses.

51. What is the sample mean difference?

52. What is the conclusion?

a. Test the above hypotheses using confidence intervals. Interpret the CI.
b. Test the above hypotheses using test statistics.
c. Test the above hypotheses using *p*-values. Interpret the *p*-value.

8.2 Comparing Two Independent Population Means

53. 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. A survey 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? Use the 1% significance level.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

54. 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. Use 95% confidence.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

55. 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 using 90% confidence.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Table 8.9.4

56. 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. Use the 5% significance level.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

57. 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. Use 99% confidence.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

Use the following information to answer the next two exercises. 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.

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

Table 8.9.5

Conduct a hypothesis test to answer the next two exercises.

58. The distribution for the hypothesis test is:

a. the normal distribution
b. the Student's *t*-distribution
c. the uniform distribution
d. the exponential distribution

59. If the level of significance is .05, the conclusion is:

a. There is sufficient evidence to conclude that the W Division teams score fewer goals, on average, than the E teams
b. There is insufficient evidence to conclude that the W Division teams score more goals, on average, than the E teams.
c. There is insufficient evidence to conclude that the W teams score fewer goals, on average, than the E teams score.
d. Unable to determine

60. 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. The "day" subscript refers to the statistics day students. The "night" subscript refers to the statistics night students. Use alpha of .05 and your choice of NHST approaches. Based on your test, which of the following would be an appropriate conclusion?

a. There is sufficient evidence to conclude that 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.

61. Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian prostitutes upon entering prostitution was 18 with a standard deviation of six. The mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

62. 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. Use 90% confidence.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

63. 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. $\mu_{\text{day}} > \mu_{\text{night}}$
b. $\mu_{\text{day}} < \mu_{\text{night}}$
c. $\mu_{\text{day}} = \mu_{\text{night}}$
d. $\mu_{\text{day}} \neq \mu_{\text{night}}$

8.4 Comparing Two Independent Population Proportions

64. 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. Use the 1% significance level to test this difference.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

65. We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for White and Black women 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. Five hundred eighty were aged 15 to 24. The estimate for Black females is 330. Forty were aged 15 to 24. Use the 5% significance level to test this difference.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

66. Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, (larger dimension/smaller dimension) = (larger dimension + smaller dimension)/(larger dimension), 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. Using 95% confidence, is there a significant difference in the Golden Ratio calculation?

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

67. 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. In general, do you think that the percentage of Hispanic students at the two colleges is basically the same or different? Use 95% confidence.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

Use the following information to answer the next two exercises. 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. 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.

68. 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.

69. An appropriate null hypothesis is:

a. $P_{2011} \leq P_{2010}$
b. $P_{2011} \geq P_{2010}$
c. $\mu_{2011} \leq \mu_{2010}$
d. $P_{2011} > P_{2010}$

70. 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. Test this research question using the 10% significance level.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

71. 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 in the table below.. Test at the 1% level of significance.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

Table 8.9.6

72. Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. Test this assertion at the 1% level of significance. The table below details the number of

tablet owners for each age group.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

	16–29 year olds	30 years old and older
Own a tablet	69	231
Sample size	628	2,309

Table 8.9.7

73. 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. Test at the 5% level of significance.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

74. While her husband spent 2½ 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. Nine of the 34 women surveyed claimed to enjoy the activity. Test this hypothesis using 95% confidence, and interpret the results of the survey.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

75. 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. Use 95% confidence to test this question.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.
e. Calculate and interpret Cohen's *d*.

76. Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is lower than the proportion of college-age females with a pierced ear. 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. Use 95% confidence here.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

8.5 Matched or Paired Samples

77. Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. The data are recorded in the table below. Were their cholesterol levels were significantly lowered? Test this question using 95% confidence.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	230

Table 8.9.8

Use the following information to answer the next two exercises. 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: $\overline{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

78. The distribution for the test is:

a. t_5
b. t_6
c. $N(-10.2, 8.4)$
d. $N\left(-10.2, \frac{8.4}{\sqrt{6}}\right)$

79. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test using 90% confidence. The data are as follows.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 8.9.9

80. 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 in the table below. Test their question using 95% confidence.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.

c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.

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

Table 8.9.10

81. 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 in the table below. Test at the 1% level of significance.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.

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	459
Philadelphia	179	159
Washington, DC	245	239

Table 8.9.11

82. A politician asked his staff to determine whether the underemployment rate in the northeast decreased from 2011 to 2012. The results are in the table below. Test with the 5% level of significance.

a. Write a pair of hypotheses to test this research question.
b. Test the above hypotheses using confidence intervals. Interpret the CI.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.

Northeastern states	2011	2012
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

Table 8.9.12

Use the following information to answer the next ten exercises. indicate which of the following choices best identifies the hypothesis test.

a. independent group means
b. matched or paired samples
c. single mean
d. two proportions
e. single proportion

83. 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.

84. 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.

85. 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 nine males and 16 females.

86. 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.

87. 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.

88. According to a YWCA Rape Crisis Center newsletter, 75% of rape victims know their attackers. A study is done to verify this.

89. According to a recent study, U.S. companies have a mean maternity-leave of six weeks.

90. A recent drug survey showed an increase in 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.

91. 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

Table 8.9.13

92. 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.

93. 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

Table 8.9.14

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

94. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four (4) new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as shown below.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 8.9.15

This is:

a. a test of two dependent means.
b. a test of two proportions.
c. a test of a single mean.
d. a test of a single proportion.

8.9: Chapter 8 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.10: Chapter 8 Solutions

1. two proportions

3. matched or paired samples

5. single mean

7. independent group means

9. two proportions

11. independent group means

13. independent group means

15. two proportions

17. The random variable is the difference between the mean amounts of sugar in the two soft drinks.

19. means

21. two-tailed

23. the difference between the mean life spans of whites and nonwhites

25. This is a comparison of two population means with unknown population standard deviations.

27. Check student's solution.

28.

1. Reject the null hypothesis
2. p-value < 0.05
3. There is not enough evidence at the 5% level of significance to support the claim that life expectancy in the 1900s is different between whites and nonwhites.

31. $P'_{OS1} - P'_{OS2}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

34. proportions

36. right-tailed

38. The random variable is the difference in proportions (percents) of the populations that are of two or more races in Nevada and North Dakota.

40. Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

42.

1. Reject the null hypothesis.
2. p-value < alpha
3. At the 5% significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

44. the mean difference of the system failures

46.

a. 99% confidence interval = [-6.17, 0.17]. Fail to reject the null hypothesis, because 0 is contained within the CI. Interpretation of CI: We're 99% confident that, after installing the software patch, there were on average between 6.17 fewer and 0.17 more system failures.
b. $t_{critical}$ = -2.998; t_{obs} = -3.31. Reject the null hypothesis, because t_{obs} exceeds $t_{critical}$.
c. In the row for 7 *df*, the closest *t*-values to our t_{obs} of -3.31 are 2.998 and 3.499, so our *p*-value must fall between .005 and .01. Interpretation of *p*-value: Assuming there truly are more or an equal number of system failures after installing the patch

compared to before installation, there is between .5% and 1% chance that we'd observe our sample mean difference of -3.

50. $H_0 : \mu_d \geq 0$; $H_a : \mu_d < 0$

52.

a. 99% confidence interval = [-3.74, 1.74]. Fail to reject the null hypothesis, because 0 is contained within the CI. Interpretation of CI: We're 99% confident that average blood pressure after 12 weeks on the medication was between 3.74 points lower and 1.74 points higher, compared to before taking the medication.
b. $t_{critical}$ = -3.365; t_{obs} = -1.47. Fail to reject the null hypothesis, because t_{obs} doesn't exceed $t_{critical}$.
c. In the row for 5 *df*, the closest *t*-value to our t_{obs} of -1.47 is (-)1.476, so our *p*-value must be greater than .10. Interpretation of *p*-value: Assuming there really was a decrease in blood pressure after 12 weeks on the medication, there is a large chance (over 10%) that we'd observe our sample mean difference of a 1 point decrease in blood pressure.

54.

a. $H_a : \mu_{4-year} > \mu_{2-year}$; $H_0 : \mu_{4-year} \leq \mu_{2-year}$
b. 95% confidence interval = [-2807.57, 3603.57]. Fail to reject the null hypothesis, because 0 is contained within the CI. Interpretation of CI: We're 95% confident that four-year colleges enroll an average of between 2807.57 fewer and 3603.57 more students than 2-year colleges do.
c. $t_{critical}$ = 1.671; t_{obs} = 0.25. Fail to reject the null hypothesis, because t_{obs} does not exceed $t_{critical}$.
d. In the row for 60 *df* (closest available in our table to 69 *df*), the closest *t*-value to our t_{obs} of 0.25 is 1.296, so our *p*-value must be more than .10. Interpretation of *p*-value: Assuming that 2-year colleges really enroll more or equal numbers of students than 4-year colleges, there is over a 10% chance that we'd observe our sample mean difference of 398 students.

56.

a. $H_a : \mu_{mechanical} < \mu_{electrical}$; $H_0 : \mu_{mechanical} \geq \mu_{electrical}$
b. 95% confidence interval = [-$2031.54, $831.54]. Fail to reject the null hypothesis, because 0 is contained within the CI. Interpretation of CI: We're 95% confident that entry-level mechanical engineers' average salary is between $2031.54 less and $831.54 more than that of entry-level electrical engineers.
c. $t_{critical}$ = -1.645; t_{obs} = -0.82. Fail to reject the null hypothesis, because t_{obs} does not exceed $t_{critical}$.
d. In the row for infinite *df*, the closest *t*-value to our t_{obs} of -0.82 is (-)1.282, so our *p*-value must be more than .10. Interpretation of *p*-value: Assuming that the mean salary of mechanical engineers is really more than or equal to that of electrical engineers, there's a good chance (over 10%) that we'd observe the data we did - a mean salary difference of $600.
e. *d* = -.16. This is a small effect.

59. c

61.

a. $H_a : \mu_{CA} < \mu_{US}$; $H_0 : \mu_{CA} \geq \mu_{US}$
b. 99% confidence interval = [-4.38, .38]. Fail to reject the null hypothesis, because 0 is contained within the CI. Interpretation of CI: We're 99% confident that the mean age of entering prostitution in Canada is between 4.38 years lower and .38 years higher than it is in the United States.
c. $t_{critical}$ = -2.326; t_{obs} = -2.166. Fail to reject the null hypothesis, because t_{obs} does not exceed $t_{critical}$.
d. In the row for infinite *df*, the closest *t*-values to our t_{obs} of -2.166 are (-)1.96 and (-)2.326, so our *p*-value must be between .01 and .025. Interpretation of *p*-value: Assuming that the mean age of entering prostitution is really higher or equal in the US compared to Canada, there is only a 1-2.5% chance that we'd observe our sample's mean difference of 2 years.
e. *d* = -.28. This is a medium-sized effect.

63. d

65.

a. $H_a : P_W \neq P_B$; $H_0 : P_W = P_B$
b. 95% confidence interval = [-.0327, .0399]. Fail to reject the null hypothesis, because 0 is contained within the CI. Interpretation of CI: We're 95% confident that the proportion of Black female suicide victims aged 15-24 is between 3.27% lower and 3.99%

higher than the proportion of White female suicide victims.

67.

a. $H_a : P_{CabrilloCollege} \neq P_{LakeTahoeCollege}$; $H_0 : P_{CabrilloCollege} = P_{LakeTahoeCollege}$
b. 95% confidence interval = [.0201, .0499]. Reject the null hypothesis, because 0 is not included in the CI. Interpretation of CI: We're 95% confident that the percentage of Hispanic students at Cabrillo College is between 2.01% and 4.99% higher than it is at Lake Tahoe College.

69. a

70.

a. $H_a : P_{16-29} \neq P_{30+}$; $H_0 : P_{16-29} = P_{30+}$
b. 90% confidence interval = [.0201, .0599]. Reject the null hypothesis, because 0 is not included in the CI. Interpretation of CI: We're 90% confident that the percentage of 30+ year olds who use eReaders is between 2.01% and 5.99% higher than the percentage of 16-29 year olds who use eReaders.

72.

a. $H_a : P_{16-29} > P_{30+}$; $H_0 : P_{16-29} \leq P_{30+}$
b. 99% confidence interval = [-.0260, .0460]. Do not reject the null hypothesis, because 0 is included in the CI. Interpretation of CI: We're 99% confident that the percentage of 16-29 year olds who use tablets is between 2.60% lower and 4.60% higher than the percentage of 30+ year olds who use tablets.

74.

a. $H_a : P_{men} > P_{women}$; $H_0 : P_{men} \leq P_{women}$
b. 95% confidence interval = [-.098, .278]. Do not reject the null hypothesis, because 0 is included in the CI. Interpretation of CI: We're 95% confident that the percentage of men who enjoy shopping for electronics is between 9.8% lower and 27.8% higher than the percentage of women who enjoy shopping for electronics.

76.

a. $H_a : P_{females} > P_{males}$; $H_0 : P_{females} \leq P_{males}$
b. 95% confidence interval = [.20, .45]. Reject the null hypothesis, because 0 is not included in the CI. Interpretation of CI: We're 95% confident that the percentage of college-age females with at least one pierced ear is between 20% and 45% more than the percentage of college-age males with at least one pierced ear.

77.

a. $H_a : \mu_{after} < \mu_{before}$; $H_0 : \mu_{after} \geq \mu_{before}$
b. 95% confidence interval = [-27.84, 7.84]. Fail to reject the null hypothesis, because 0 is contained within the CI. Interpretation of CI: We're 95% confident that the mean cholesterol level after dieting is between 27.84 points lower and 7.84 points higher than it was beforehand.
c. $t_{critical}$ = -1.833; t_{obs} = -1.27. Fail to reject the null hypothesis, because t_{obs} does not exceed $t_{critical}$.
d. In the row for 9 *df*, the closest *t*-value to our t_{obs} of -1.27 is (-)1.383, so our *p*-value must be more than .10. Interpretation of *p*-value: Assuming that cholesterol levels really increase or remain the same on this diet, there's more than a 10% chance that we'd get our sample mean change in cholesterol levels of -10.

80.

a. $H_a : \mu_{2013} > \mu_{2012}$; $H_0 : \mu_{2013} \leq \mu_{2012}$
b. 95% confidence interval = [88.45, 261.55]. We can reject the null hypothesis, because 0 isn't contained within the CI. Interpretation of CI: We're 95% confident that there were an average of 88.45 and 261.55 more new female breast cancer cases in the south in 2013, compared to 2012.
c. $t_{critical}$ = 1.782; t_{obs} = 4.41. Reject the null hypothesis, because t_{obs} exceeds $t_{critical}$.
d. In the row for 12 *df*, the closest *t*-value to our t_{obs} of 4.41 is 3.930, so our *p*-value must be less than .001. Interpretation of *p*-value: Assuming that the number of new female breast cancer cases really increased or remained equal from 2012 to 2013,

there's less than a 0.1% chance that we'd observe our sample mean increase in cases of 175.

81.

a. $H_a : \mu_{Hyatt} \neq \mu_{Hilton}$; $H_0 : \mu_{Hyatt} = \mu_{Hilton}$

b. 99% confidence interval = [-$81.90, $63.90]. We cannot reject the null hypothesis, because 0 is included within the CI. Interpretation of CI: We're 99% confident that Hilton prices, on average, are between $81.90 lower and $63.90 higher than Hyatt prices in the same cities.

c. $t_{critical}$ = ±3.250; t_{obs} = -.40. Do not reject the null hypothesis, because t_{obs} does not exceed $t_{critical}$.

d. In the row for 9 *df*, the closest *t*-value to our t_{obs} of -.40 is (-)1.383, so our *p*-value must be more than .10 x 2 (because we have a two-tailed hypothesis), or .20. Interpretation of *p*-value: Assuming that Hilton and Hyatt prices (within in the same cities) truly do not differ from each other, there's over a 20% chance that we'd observe our sample mean price difference of $9.

84. d

86. c

88. e

90. d

92. e

94. a

8.10: Chapter 8 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **8.9: Chapter 8 Homework** has no license indicated.

8.11: Chapter 8 References

8.2 Comparing Two Independent Population Means

Data from Graduating Engineer + Computer Careers. Available online at http://www.graduatingengineer.com

Data from *Microsoft Bookshelf*.

Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).

"List of current United States Senators by Age." Wikipedia. Available online at http://en.Wikipedia.org/wiki/List_of...enators_by_age (accessed June 17, 2013).

"Sectoring by Industry Groups." Nasdaq. Available online at http://www.nasdaq.com/markets/barcha...&base=industry (accessed June 17, 2013).

"Strip Clubs: Where Prostitution and Trafficking Happen." Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).

"World Series History." Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).

8.4 Comparing Two Independent Population Proportions

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/St...ef=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).

8.11: Chapter 8 References is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

9: F-Distribution and One-Way ANOVA

9.1: Introduction

9.2: One-Way ANOVA

9.3: The F-Distribution and the F-Ratio

9.4: Chapter 9 Key Terms

9.5: Chapter 9 Review

9.6: Chapter 9 Formula Review

9.7: Chapter 9 Homework

9.8: Chapter 9 Solutions

9.9: Chapter 9 References

This page titled 9: F-Distribution and One-Way ANOVA is shared under a CC BY license and was authored, remixed, and/or curated by .

9.1: Introduction



Figure 9.1.1 One-way ANOVA is used to measure information from several groups.

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests comparing averages among more than two groups, statisticians have developed a method called "Analysis of Variance" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the F distribution, used for one-way ANOVA.

This page titled 9.1: Introduction is shared under a CC BY license and was authored, remixed, and/or curated by .

9.2: One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test actually uses **variances** to help determine if the means are equal or not. In order to perform a one-way ANOVA test, there are five basic **assumptions** to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have **equal standard deviations (or variances)**.
4. The factor is a categorical variable.
5. The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are *g* groups:

$H_0 : \mu_1 = \mu_2 = \mu_3 = ... \mu_g$

H_a: At least two of the group means $\mu_1, \mu_2, \mu_3, ... , \mu_g$ are not equal. That is, $\mu_i \neq \mu_j$ for some $i \neq j$.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), $H_0 : \mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).

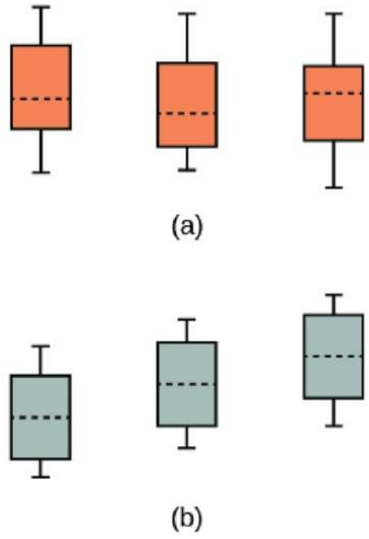


Figure 9.2.1 (a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.

This page titled 9.2: One-Way ANOVA is shared under a CC BY license and was authored, remixed, and/or curated by .

9.3: The F-Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the ***F*-distribution**, invented by George Snedecor but named in honor of Sir Ronald Fisher, an English statistician. The F-statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

Here are some facts about the *F*-distribution.

1. The curve is not symmetrical but skewed to the right.
2. There is a different curve for each set of degrees of freedom.
3. The F-statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal as can be seen in the two figures below. Notice that with more degrees of freedom as shown in the figure below, the curve is more closely approaching the normal distribution, but remember that the F cannot ever be less than zero so the distribution does not have a tail that goes to infinity on the left as the normal distribution does.
5. Other uses for the F-distribution include comparing two variances and two-way Analysis of Variance. Two-Way Analysis is beyond the scope of this chapter.

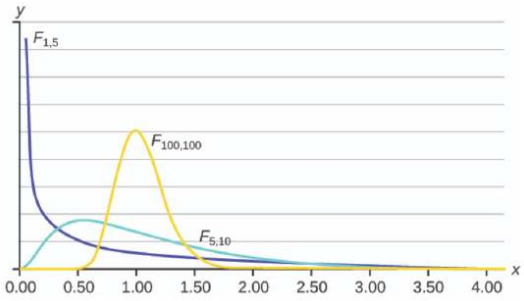


Figure 9.3.1

For example, if F follows an F distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

To calculate the ***F*-ratio**, two estimates of the variance are made.

1. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by *n*. (If the samples are different sizes, the variance between samples is thus weighted to account for the different sample sizes.) This variance is also called **variation due to treatment or group,** or **explained variation.** It may also be called **"factor" variance**.

$$MS_{\text{Between}} = \frac{SS_{\text{Between}}}{df_{\text{Between}}} = \frac{n_1 (\bar{x}_1 - \bar{x})^2 + \cdots + n_g (\bar{x}_g - \bar{x})^2}{g - 1}$$

2. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances. When the sample sizes are different, the variance within samples is weighted. This variance is also called the **variation due to error,** or **unexplained variation.**

$$MS_{\text{Within}} = \frac{SS_{\text{Within}}}{df_{\text{Within}}} = \frac{(n_1 - 1) s_1^2 + \cdots + (n_g - 1) s_g^2}{n - g}$$

- $SS_{between}$ = the **sum of squares** that represents the variation among the different samples
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation.

MS in these equations means "**mean square**". $MS_{between}$ is the variance between groups, and MS_{within} is the variance within groups.

The one-way ANOVA test depends on the fact that $MS_{between}$ can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different

does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions.

Note

The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution, because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or *F*-Statistic

$$F_{obs} = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

If $MS_{between}$ and MS_{within} estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, $MS_{between}$ consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. If the null hypothesis is false, $MS_{between}$ will generally be larger than MS_{within}, and the F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

To determine the critical value, we have to find $F_{\alpha,df1,df2}$. See Appendix A for the F table. This table has F values for various levels of significance on different pages, as indicated in the first row of each page's table. To find the critical value, choose the page/table with the desired significance level, and follow down and across to find the critical value at the intersection of the two different degrees of freedom. The F distribution has two different degrees of freedom, one associated with the numerator, df_1, and one associated with the denominator, df_2. The degrees of freedom in the numerator is $g - 1$, where g is the number of groups; and the degrees of freedom in the denominator is $n - g$, where n is the total sample size across all groups. $F_{\alpha,df1,df2}$ will give the critical value on the upper end of the F distribution.

Data are typically put into a table for easy viewing. One-way ANOVA results are often displayed in this manner by computer software.

Table 9.3.1

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F
Factor (Between)	SS(Factor)	g–1	$MS(Factor) = SS(Factor)/(g–1)$	$F = MS(Factor)/MS(Error)$
Error (Within)	SS(Error)	$n–g$	$MS(Error) = SS(Error)/(n–g)$	
Total	SS(Total)	n–1		

Example 9.3.1

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in the table below.

Table 9.3.2

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4	7	4
4	4.5	3
3		

$H_0 : \mu_1 = \mu_2 = \mu_3$
H_a: At least two of the means (μ_1, μ_2, μ_3) are not equal.
Means: $\bar{x}_1 = \frac{16}{4} = 4, \bar{x}_2 = \frac{15}{3} = 5, \bar{x}_3 = \frac{15}{3} = 5, \bar{x} = \frac{46}{10} = 4.6$
Variances: $s_1^2 = 0.67, s_2^2 = 3.16, s_3^2 = 7$
$SS_{Between} = 4(4 - 4.6)^2 + 3(5 - 4.6)^2 + 3(5 - 4.6)^2 = 2.4$
$SS_{Within} = (4 - 1)0.67 + (3 - 1)3.16 + (3 - 1)7 = 22.33$
$df_{Between} = 3 - 1 = 2$
$df_{Within} = 10 - 3 = 7$
$MS_{Between} = \frac{SS_{Between}}{df_{Between}} = \frac{2.4}{2} = 1.2$
$MS_{Within} = \frac{SS_{Within}}{df_{Within}} = \frac{22.33}{7} = 3.19$
$F_{obs} = \frac{MS_{Between}}{MS_{Within}} = \frac{1.2}{3.19} = 0.38$

Table 9.3.3

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F
Factor (Between)	2.4	2	1.2	0.38
Error (Within)	22.33	7	3.19	
Total	2.4 + 22.33 = 24.73	2 + 7 = 9		

For $F_{\alpha=.05,2,7}$, $F_{critical}$ = 4.74. F_{obs} does not exceed this $F_{critical}$, so we cannot reject H_0; we don't have sufficient evidence to say that the three different diet plans result in different average amounts of weight loss.

Exercise 9.3.1

As part of an experiment to see how different types of soil cover would affect slicing tomato production, Marist College students grew tomato plants under different soil cover conditions. Groups of three plants each had one of the following treatments

- bare soil
- a commercial ground cover
- black plastic
- straw
- compost

All plants grew under the same conditions and were the same variety. Students recorded the weight (in grams) of tomatoes produced by each of the *n* = 15 plants:

Table 9.3.4

Bare: $n_1 = 3$	Ground Cover: $n_2 = 3$	Plastic: $n_3 = 3$	Straw: $n_4 = 3$	Compost: $n_5 = 3$
2,625	5,348	6,583	7,285	6,277
2,997	5,682	8,560	6,897	7,818
4,915	5,482	3,830	9,230	8,677

Create the one-way ANOVA table.

The one-way ANOVA hypothesis test is always right-tailed because larger F-values are way out in the right tail of the F-distribution curve.

Example 9.3.2

Let's return to the slicing tomato exercise we were working on in the previous exercise. The means of the tomato yields under the five mulching conditions are represented by $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$. We will conduct a hypothesis test to determine if all means are the same or at least one is different. Using a significance level of 5%, test the null hypothesis that there is no difference in mean yields among the five groups against the alternative hypothesis that at least one mean is different from the rest.

Answer

The null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a : at least two of the population means differ

The one-way ANOVA results are shown in the table below.

Table 9.3.5

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F
Factor (Between)	36,648,561	5–1 = 4	$\frac{36,648,561}{4} = 9,162,140$	$\frac{9,162,140}{2,044,672.6} = 4.481$
Error (Within)	20,446,726	15–5 = 10	$\frac{20,446,726}{10} = 2,044,672.6$	
Total	57,095,287	15–1 = 14		

Distribution for the test: $F_{\alpha=.05,4,10}$

$df_{num} = 5–1 = 4$

$df_{denom} = 15–5 = 10$

$F_{critical} = 3.48$

Test statistic: $F_{obs} = 4.481$

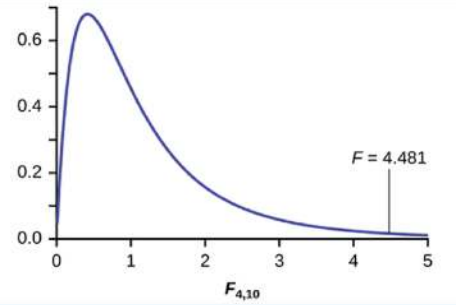


Figure 9.3.2

Make a decision: Since $F_{obs} > F_{critical}$, we reject H_0.

Conclusion: At the 5% significance level, we have reasonably strong evidence that differences in mean yields for slicing tomato plants grown under different mulching conditions are unlikely to be due to chance alone. We may conclude that at least one of mulches led to a different mean yield than the others.

Exercise 9.3.2

MRSA, or *Staphylococcus aureus*, can cause a serious bacterial infections in hospital patients. The table below shows various colony counts from different patients who may or may not have MRSA. The data from the table is plotted in the figure below.

Table 9.3.6

Conc = 0.6	Conc = 0.8	Conc = 1.0	Conc = 1.2	Conc = 1.4
9	16	22	30	27
66	93	147	199	168
98	82	120	148	132

Plot of the data for the different concentrations:

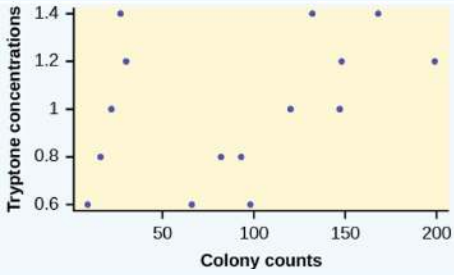


Figure 9.3.3

Test whether the mean number of colonies are the same or are different. Construct the ANOVA table, find the *p*-value, and state your conclusion. Use a 5% significance level.

Example 9.3.3

Four sororities took a random sample of their members regarding their grade means for the past term. The results are shown in the table below.

Table 9.3.7

Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Using a significance level of 1%, is there a difference in mean grades among the sororities?

Answer

Note

This is an example of a **balanced design**, because each group (i.e., sorority) has the same number of observations.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means $\mu_1, \mu_2, \mu_3, \mu_4$ are equal.

Distribution for the test: $F_{\alpha=.01,3,16}$

where $g = 4$ groups and $n = 20$ samples in total

$df_{num} = g–1 = 4–1 = 3$

$df_{denom} = n–g = 20–4 = 16$

$F_{critical} = 5.29$

Calculate the test statistic: $F_{obs} = 2.23$

Graph:

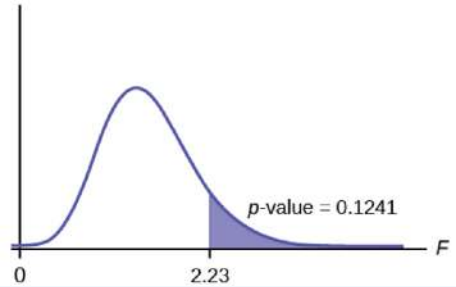


Figure 9.3.4

Make a decision: Since $F_{obs} < F_{critical}$, we reject H_0.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

Exercise 9.3.3

Four sports teams took a random sample of players regarding their GPAs for the last year. The results are shown here in the table.

Table 9.3.8

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Use a significance level of 5%, and determine if there is a difference in GPA among the teams.

Example 9.3.4

A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data (in inches) in the table shown here.

Table 9.3.9

Tommy's plants	Tara's plants	Nick's plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Does it appear that the three media in which the bean plants were grown produce the same mean height? Test at a 5% level of significance.

Answer

$H_0 : \mu_{Tommy} = \mu_{Tara} = \mu_{Nick}$
H_a: At least two of the means ($\mu_{Tommy}, \mu_{Tara}, \mu_{Nick}$) are not equal.
Means: $\bar{x}_{Tommy} = \frac{121}{5} = 24.2, \bar{x}_{Tara} = \frac{127}{5} = 25.4, \bar{x}_{Nick} = \frac{122}{5} = 24.4, \bar{x} = \frac{370}{15} = 24.7$
Variances: $s^2_{Tommy} = 11.7, s^2_{Tara} = 18.3, s^2_{Nick} = 16.3$
$SS_{Between} = 5(24.2 - 24.7)^2 + 5(25.4 - 24.7)^2 + 5(24.4 - 24.7)^2 = 4.15$
$SS_{Within} = (5 - 1)11.7 + (5 - 1)18.3 + (5 - 1)16.3 = 185.2$
$df_{Between} = 3 - 1 = 2$
$df_{Within} = 15 - 3 = 12$
$MS_{Between} = \frac{SS_{Between}}{df_{Between}} = \frac{4.15}{2} = 2.075$
$MS_{Within} = \frac{SS_{Within}}{df_{Within}} = \frac{185.2}{12} = 15.43$
$F_{obs} = \frac{MS_{Between}}{MS_{Within}} = \frac{2.075}{15.43} = 0.13$

Table 9.3.10

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F
Factor (Between)	4.15	2	2.075	0.13
Error (Within)	185.2	12	15.43	
Total	189.35	14		

For $F_{\alpha=.05,2,12}$, $F_{critical}$ = 3.89. F_{obs} does not exceed this $F_{critical}$, so we cannot reject H_0. With a 5% level of significance, from this sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

This page titled 9.3: The F-Distribution and the F-Ratio is shared under a CC BY license and was authored, remixed, and/or curated by .

9.4: Chapter 9 Key Terms

Analysis of Variance

also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

Between-Groups Variance

an estimate of the average variance among sample means (i.e., from the different groups) across all of the groups.

One-Way ANOVA

an analysis of variance with *one* independent (grouping) variable.

Within-Groups Variance

an estimate of the average of the sample variances (i.e., within the different groups) across all of the groups.

This page titled 9.4: Chapter 9 Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by .

9.5: Chapter 9 Review

9.2 One-Way ANOVA

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent, and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the F distribution with two different degrees of freedom.

Assumptions:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).
4. The factor is a categorical variable.
5. The response is a numerical variable.

9.3 The *F*-Distribution and the *F*-Ratio

Analysis of variance compares the means of a response variable for several groups. ANOVA compares the variation within each group (within-groups variance) to the variation among the means of each group (between-groups variance). The ratio of these two is the F-statistic from an F-distribution with (number of groups – 1) as the numerator degrees of freedom and (number of observations – number of groups) as the denominator degrees of freedom. These statistics are summarized in the ANOVA table.

The graph of the F-distribution is always positive and skewed right, though the shape can be mounded or exponential depending on the combination of numerator and denominator degrees of freedom. The F-statistic is the ratio of a measure of the variation in the group means to a similar measure of the variation within the groups. If the null hypothesis is correct, then the numerator should be small compared to the denominator. A small F-statistic will result, and the area under the F-curve to the right will be large, representing a large p-value. When the null hypothesis of equal group means is incorrect, then the numerator should be large compared to the denominator, giving a large F-statistic and a small area (small p-value) to the right of the statistic under the F-curve.

9.5: Chapter 9 Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.6: Chapter 9 Formula Review

9.3 The *F*-Distribution and the *F*-Ratio

$$MS_{\text{Between}} = \frac{SS_{\text{Between}}}{df_{\text{Between}}} = \frac{n_1 (\bar{x}_1 - \bar{x})^2 + \cdots + n_g (\bar{x}_g - \bar{x})^2}{g - 1}$$

$$MS_{\text{Within}} = \frac{SS_{\text{Within}}}{df_{\text{Within}}} = \frac{(n_1 - 1) s_1^2 + \cdots + (n_g - 1) s_g^2}{n - g}$$

$$F_{obs} = \frac{MS_{\text{ between}}}{MS_{\text{ within}}}$$

- n_g = the size of the *g*th group
- $\bar{x}_g$ = the mean of the values in the *g*th group
- $\bar{x}$ = the total mean of all observations
- g = the number of groups
- n = the total number of observations combined
- s_g^2 = the variance of the values in the *g*th group

9.6: Chapter 9 Formula Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.7: Chapter 9 Homework

9.3 The *F*-Distribution and the *F*-Ratio

1. There are five basic assumptions that must be fulfilled in order to perform a one-way ANOVA test. What are they?

Use the following information to answer the next eleven exercises. Groups of men from three different areas of the country are to be tested for mean weight. The entries in Table 9.7.1 are the weights for the different groups.

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

Table 9.7.1

2. State the null and alternative hypotheses.

3. What is the Sum of Squares Factor?

4. What is the Sum of Squares Error?

5. What is the df for the numerator?

6. What is the df for the denominator?

7. What is the Mean Square Factor?

8. What is the Mean Square Error?

9. What is the F-observed value?

10. Create an ANOVA summary table.

11. What is the F-critical value for the 95% confidence level? Make a decision about the hypothesis.

12. What is the approximate p-value for F-observed here? Make a decision about the hypothesis using $\alpha = .05$.

Use the following information to answer the next eleven exercises. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in Table 9.7.2 are the goals per game for the different teams.

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

Table 9.7.2

13. State the null and alternative hypotheses.

14. What is $SS_{Between}$?

15. What is the df for the numerator?

16. What is $MS_{Between}$?

17. What is SS_{Within}?

18. What is the df for the denominator?

19. What is MS_{Within}?

20. What is the F-observed value?

21. Create an ANOVA summary table.

22. What is the F-critical value for the 95% confidence level? Make a decision about the hypothesis.

23. What is the approximate p-value for F-observed here? Make a decision about the hypothesis using α = .05.

24. An F-statistic can have what values?

25. What happens to the curves as the degrees of freedom for the numerator and the denominator get larger?

Use the following information to answer the next ten exercises. Four basketball teams took a random sample of players regarding how high each player can jump (in inches). The results are shown in Table 9.7.3.

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

Table 9.7.3

26. State the null and alternative hypotheses.

27. What is the df_{num}?

28. What is the df_{denom}?

29. What are the Sum of Squares and Mean Squares Factors?

30. What are the Sum of Squares and Mean Squares Errors?

31. What is the F-observed statistic?

32. Create an ANOVA summary table.

33. What is the F-critical value for the 95% confidence level?

34. What is the approximate p-value for F-observed here?

35. At the 5% significance level, is there a difference in the mean jump heights among the teams?

Use the following information to answer the next ten exercises. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are shown in Table 9.7.4.

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

Table 9.7.4

36. State the null and alternative hypotheses.

37. What is the df_{num}?

38. What is the df_{denom}?

39. What are the $SS_{Between}$ and $MS_{Between}$?

40. What are the SS_{Within} and MS_{Within}?

41. What is the F-observed statistic?

42. Create an ANOVA summary table.

43. What is the F-critical value for the 99% confidence level? Make a decision about the hypothesis.

44. What is the approximate p-value for F-observed here? Make a decision about the hypothesis using α = .01.

45. At the 1% significance level, are the scores among the different groups different? Why or why not?

Use the following information to answer the next nine exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from four teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x}$=	_____	_____	_____	_____	_____
s^2=	_____	_____	_____	_____	_____

Table 9.7.5

46. State the hypotheses.

H_0: _____

H_a: _____

47. sum of squares – between groups: $SS_{Between}$ = _____

48. sum of squares – within groups: SS_{Within} = _____

49. degrees of freedom – numerator: df_{num} = _____

50. degrees of freedom – denominator: df_{denom} = _____

51. F-observed = _____

52. Approximate p-value = _____

State the decisions and conclusions (in complete sentences) for the following preconceived levels of α.

53. $\alpha = 0.05$

a. Decision: _____

b. Conclusion: _____

54. $\alpha = 0.01$

a. Decision: _____

b. Conclusion: _____

55. The American League and the National League of Major League Baseball are each divided into three divisions: East, Central, and West. Many years, fans talk about some divisions being stronger (having better teams) than other divisions. This may have consequences for the postseason. For instance, in 2012 Tampa Bay won 90 games and did not play in the postseason, while Detroit won only 88 and did play in the postseason. This may have been an oddity, but is there good evidence that in the 2012 season, the American League divisions were significantly different in overall records? Use the following data to complete the test statistic approach whether the mean number of wins per team in the three American League divisions were the same or not. Note that the data are not balanced, as two divisions had five teams, while one had only four. Use α = .05.

Division	Team	Wins
East	NY Yankees	95
East	Baltimore	93
East	Tampa Bay	90
East	Toronto	73
East	Boston	69

Table 9.7.6

Division	Team	Wins
Central	Detroit	88
Central	Chicago Sox	85
Central	Kansas City	72
Central	Cleveland	68
Central	Minnesota	66

Table 9.7.7

Division	Team	Wins
West	Oakland	94
West	Texas	93
West	LA Angels	89
West	Seattle	75

Table 9.7.8

9.2 One-Way ANOVA

56. Three different traffic routes are tested for mean driving time. The entries in Table 9.7.9 are the driving times in minutes on the three different routes. Conduct a hypothesis test using the test statistic approach and 5% significance, and create an ANOVA summary table of the results.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

Table 9.7.9

9.3 The *F*-Distribution and the F-Ratio

57. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 1% and the test statistic approach, test the hypothesis that the three formulas produce the same mean weight gain.

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Table 9.7.10 Weights of Student Lab Rats

58. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are in Table 9.7.11. Using a 5% significance level and the test statistic approach, test the hypothesis that the three mean commuting mileages are the same.

Working-class	Professional (middle incomes)	Professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Table 9.7.11

Use the following information to answer the next two exercises. Table 9.7.12 lists the number of pages in four different types of magazines.

Home decorating	News	Health	Computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

Table 9.7.12

59. Using a significance level of 5%, use test statistics test whether the four magazine types have the same mean length. Create an ANOVA summary table of the results.

60. Eliminate one magazine type that you now feel has a mean length different from the others. Redo the hypothesis test, testing that the remaining three means are statistically the same. Create a new ANOVA summary table. Based on this test, are the mean lengths for the remaining three magazines statistically the same?

61. A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. Suppose that Table 9.7.13 shows the results of a study. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use $\alpha = 0.05$ and the test statistic approach.

CNN	FOX	Local
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

Table 9.7.13

62. Are the means for the final exams the same for all statistics class delivery types? Table 9.7.14 shows the scores on final exams from several randomly selected classes that used the different delivery types. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05 to test this with test statistics.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

Table 9.7.14

63. Are the mean number of times a month a person eats out the same for Whites, Blacks, Hispanics and Asians? Suppose that Table 9.7.15 shows the results of a study. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.01 and the test statistic approach. Create an ANOVA summary table.

White	Black	Hispanic	Asian
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

Table 9.7.15

64. Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? Suppose that Table 9.7.16 shows the results of a study. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05 and the test statistic approach.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

Table 9.7.16

65. Sanjay made identical paper airplanes out of three different weights of paper, light, medium and heavy. He made four airplanes from each of the weights, and launched them himself across the room. Here are the distances (in meters) that his planes flew.

Paper type/Trial	Trial 1	Trial 2	Trial 3	Trial 4
Heavy	5.1 meters	3.1 meters	4.7 meters	5.3 meters
Medium	4 meters	3.5 meters	4.5 meters	6.1 meters
Light	3.1 meters	3.3 meters	2.1 meters	1.9 meters

Table 9.7.17

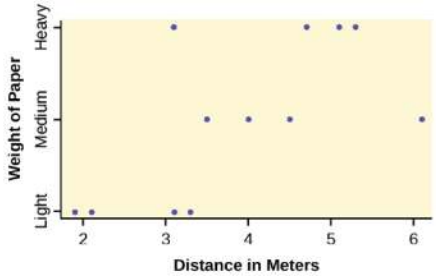


Figure 9.7.1

1. Take a look at the data in the graph. Look at the spread of data for each group (light, medium, heavy). Does it seem reasonable to assume a normal distribution with the same variance for each group? Yes or No.
2. Why is this a balanced design?
3. Calculate the sample mean and sample standard deviation for each group.
4. Does the weight of the paper have an effect on how far the plane will travel? Use a 1% level of significance.
   - g = _____
   - n = _____
   - $SS_{Between}$= _____
   - $MS_{Between}$= _____
   - SS_{Within}= _____
   - MS_{Within} = _____
   - df_{num} = _____, df_{denom} = _____
   - F statistic = _____
   - F-critical = _____
   - Graph the F-observed and F-critical values.
   - decision: _____
   - conclusion: _____

66. DDT is a pesticide that has been banned from use in the United States and most other areas of the world. It is quite effective, but persisted in the environment and over time became seen as harmful to higher-level organisms. Famously, egg shells of eagles and other raptors were believed to be thinner and prone to breakage in the nest because of ingestion of DDT in the food chain of the birds.

An experiment was conducted on the number of eggs (fecundity) laid by female fruit flies. There are three groups of flies. One group was bred to be resistant to DDT (the RS group). Another was bred to be especially susceptible to DDT (SS). Finally there was a control line of non-selected or typical fruitflies (NS). Here are the data:

RS	SS	NS	RS	SS	NS
12.8	38.4	35.4	22.4	23.1	22.6
21.6	32.9	27.4	27.5	29.4	40.4
14.8	48.5	19.3	20.3	16	34.4
23.1	20.9	41.8	38.7	20.1	30.4
34.6	11.6	20.3	26.4	23.3	14.9
19.7	22.3	37.6	23.7	22.9	51.8
22.6	30.2	36.9	26.1	22.5	33.8
29.6	33.4	37.3	29.5	15.1	37.9
16.4	26.7	28.2	38.6	31	29.5
20.3	39	23.4	44.4	16.9	42.4
29.3	12.8	33.7	23.2	16.1	36.6
14.9	14.6	29.2	23.6	10.8	47.4
27.3	12.2	41.7			

Table 9.7.18

The values are the average number of eggs laid daily for each of 75 flies (25 in each group) over the first 14 days of their lives. Using a 1% level of significance and the *p*-value approach, are the mean rates of egg selection for the three strains of fruitfly different? If so, in what way? Specifically, the researchers were interested in whether or not the selectively bred strains were different from the non-selected line, and whether the two selected lines were different from each other.

Here is a chart of the three groups:

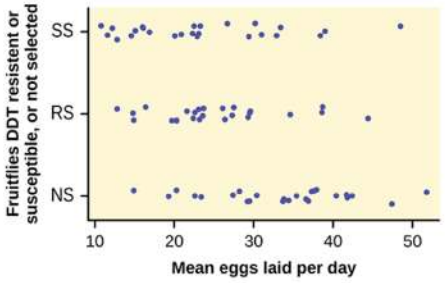


Figure 9.7.2

67. The data shown is the recorded body temperatures of 130 subjects as estimated from available histograms.

Traditionally we are taught that the normal human body temperature is 98.6 F. This is not quite correct for everyone. Are the mean temperatures among the four groups different? Conduct both the test statistic and *p*-value approach using 99% confidence.

FL	FH	ML	MH	FL	FH	ML	MH
96.4	96.8	96.3	96.9	98.4	98.6	98.1	98.6
96.7	97.7	96.7	97	98.7	98.6	98.1	98.6
97.2	97.8	97.1	97.1	98.7	98.6	98.2	98.7
97.2	97.9	97.2	97.1	98.7	98.7	98.2	98.8
97.4	98	97.3	97.4	98.7	98.7	98.2	98.8
97.6	98	97.4	97.5	98.8	98.8	98.2	98.8
97.7	98	97.4	97.6	98.8	98.8	98.3	98.9
97.8	98	97.4	97.7	98.8	98.8	98.4	99
97.8	98.1	97.5	97.8	98.8	98.9	98.4	99
97.9	98.3	97.6	97.9	99.2	99	98.5	99
97.9	98.3	97.6	98	99.3	99	98.5	99.2
98	98.3	97.8	98		99.1	98.6	99.5
98.2	98.4	97.8	98		99.1	98.6	
98.2	98.4	97.8	98.3		99.2	98.7	
98.2	98.4	97.9	98.4		99.4	99.1	
98.2	98.4	98	98.4		99.9	99.3	
98.2	98.5	98	98.6		100	99.4	
98.2	98.6	98	98.6		100.8		

Table 9.7.19

9.7: Chapter 9 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.8: Chapter 9 Solutions

3. $SS_{Factor} = 4,939.2$

5. $df_{num} = 2$

7. $MS_{Factor} = 2,469.6$

9. $F_{obs} = 3.742$

11. $F_{\alpha=.05,2,12} = 3.89$. F_{obs} does not exceed $F_{critical}$, so we fail to reject H_0.

13. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$; H_a: At least two of the means $(\mu_1, \mu_2, \mu_3, \mu_4)$ are not equal.

15. $df_{num} = 3$

17. $SS_{Within} = 13.2$

19. $MS_{Within} = .825$

21.

Source of variation	SS	df	MS	F
Between	25.75	3	8.583	10.404
Within	13.2	16	.825	
Total	38.95	19		

Table 9.8.1

23. The closest available F-value to our $F_{obs} = 10.404$ at (3, 16) df is 9.01, which corresponds to α of .01. Therefore, the p-value must be less than .01. Because this p-value is less than α = .05, we reject H_0.

27. $df_{num} = 4$

29. $SS_{Factor} = 195.6$; $MS_{Factor} = 48.9$

31. $F_{obs} = 2.060$

33. $F_{\alpha=.05,4,10} = 3.48$

35. No, F_{obs} does not exceed $F_{critical}$, and the approximate p-value exceeds α, so we fail to reject H_0.

37. $df_{num} = 2$

39. $SS_{Between} = 5,700.4$; $MS_{Between} = 2,850.2$

41. $F_{obs} = 3.610$

43. $F_{\alpha=.01,2,12} = 6.93$

45. No, F_{obs} does not exceed $F_{critical}$, and the approximate p-value exceeds α, so we fail to reject H_0.

47. $SS_{Between} = .903$

49. $df_{num} = 4$

51. $F_{obs} = 4.220$

53.

a. Decision: Reject H_0, because the approximate p-value (between .01 and .05) is less than α.

b. Conclusion: There is a significant difference in at least one of these regions of the country for the average age at which teenagers obtain their driver's licenses.

55. $H_0 : \mu_{East} = \mu_{Central} = \mu_{West}$; H_a: At least two of the means $(\mu_{East}, \mu_{Central}, \mu_{West})$ are not equal.

$F_{obs} = \frac{\frac{344.164}{2}}{\frac{1,219.550}{11}} = \frac{172.082}{110.868} = 1.552$

$F_{\alpha=.05,2,12} = 3.98$

F_{obs} does not exceed $F_{critical}$, so we fail to reject H_0.

57. $H_0 : \mu_{Linda} = \mu_{Tuan} = \mu_{Javier}$; H_a: At least two of the means ($\mu_{Linda}, \mu_{Tuan}, \mu_{Javier}$) are not equal.

$F_{obs} = \frac{\frac{23.212}{2}}{\frac{208.324}{12}} = \frac{11.606}{17.360} = .669$

$F_{\alpha=.01,2,12} = 6.93$

F_{obs} does not exceed $F_{critical}$, so we fail to reject H_0.

59. $H_0 : \mu_{Home} = \mu_{News} = \mu_{Health} = \mu_{Computer}$; H_a: At least two of the means ($\mu_{Home}, \mu_{News}, \mu_{Health}, \mu_{Computers}$) are not equal.

Source of variation	SS	df	MS	F
Between	34,288.6	3	11,429.533	8.689
Within	21,047.6	16	1,315.475	
Total	55,336.2	19		

Table 9.8.2

$F_{\alpha=.05,3,16} = 3.24$. F_{obs} does exceed $F_{critical}$, so we can reject H_0.

61. $H_0 : \mu_{CNN} = \mu_{FOX} = \mu_{Local}$; H_a: At least two of the means ($\mu_{CNN}, \mu_{FOX}, \mu_{Local}$) are not equal.

$F_{obs} = \frac{\frac{1,967.925}{2}}{\frac{3,375.133}{14}} = \frac{983.963}{241.081} = 4.081$

$F_{\alpha=.05,2,14} = 3.74$

F_{obs} exceeds $F_{critical}$, so we reject H_0.

63. $H_0 : \mu_{White} = \mu_{Black} = \mu_{Hispanic} = \mu_{Asian}$; H_a: At least two of the means ($\mu_{White}, \mu_{Black}, \mu_{Hispanic}, \mu_{Asian}$) are not equal.

Source of variation	SS	df	MS	F
Between	13.032	3	4.344	.885
Within	73.600	15	4.907	
Total	86.632	18		

Table 9.8.3

$F_{\alpha=.01,3,15} = 5.42$. F_{obs} doesn't exceed $F_{critical}$, so we cannot reject H_0.

65.

1. Yes.
2. Because $n_1 = n_2 = n_3$.
3. $\bar{x}_{Heavy} = 4.55, s_{Heavy} = 1.00; \bar{x}_{Medium} = 4.525, s_{Medium} = 1.13; \bar{x}_{Light} = 2.60, s_{Light} = .70$
4. - g = 3
   - n = 12
   - $SS_{Between}$ = 10.012
   - $MS_{Between}$ = 5.006
   - SS_{Within} = 8.277
   - MS_{Within} = .920
   - df_{num} = 2, df_{denom} = 9
   - F statistic = 5.443

- F-critical = 8.02
- Graph: Check student's solution.
- Decision: F_{obs} doesn't exceed $F_{critical}$, so we cannot reject H_0.
- Conclusion: There is not sufficient evidence to conclude that the mean paper airplane flight distances differ based on paper weights.

9.8: Chapter 9 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **9.7: Chapter 9 Homework** has no license indicated.

9.9: Chapter 9 References

9.3 The *F*-Distribution and the *F*-Ratio

Tomato Data, Marist College School of Science (unpublished student research)

Data from a fourth grade classroom in 1994 in a private K – 12 school in San Jose, CA.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets:*. London: Chapman & Hall, 1994.

“MLB Standings – 2012.” Available online at http://espn.go.com/mlb/standings/_/year/2012.

Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, 268, 1578-1580.

9.9: Chapter 9 References is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

10: Linear Regression and Correlation

10.1: Introduction

10.2: The Correlation Coefficient r

10.3: Testing the Significance of the Correlation Coefficient

10.4: Linear Equations

10.5: The Regression Equation

10.6: How to Use Microsoft Excel® for Regression Analysis

10.7: Chapter 10 Key Terms

10.8: Chapter 10 Review

10.9: Chapter 10 Homework

10.10: Chapter 10 Solutions

This page titled 10: Linear Regression and Correlation is shared under a CC BY license and was authored, remixed, and/or curated by .

10.1: Introduction



Figure 10.1.1 Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

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

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

These examples may or may not be tied to a model, meaning that some theory suggested that a relationship exists. This link between a cause and an effect, often referred to as a model, is the foundation of the scientific method and is the core of how we determine what we believe about how the world works. Beginning with a theory and developing a model of the theoretical relationship should result in a prediction, what we have called a hypothesis earlier. Now the hypothesis concerns a full set of relationships. As an example, in economics the model of consumer choice is based upon assumptions concerning human behavior: a desire to maximize something called utility, knowledge about the benefits of one product over another, likes and dislikes, referred to generally as preferences, and so on. These are combined to give us the demand curve. From that we have the prediction that as prices rise the quantity demanded will fall. Economics has models concerning the relationship between what prices are charged for goods and the market structure in which the firm operates, monopoly versus perfect competition, for example. We can generate models for who would most likely be chosen for an on-the-job training position, the impacts of Federal Reserve policy changes on the growth of the economy and so on.

Models are not unique to economics, even within the social sciences. In political science, for example, there are models that predict behavior of bureaucrats to various changes in circumstances based upon assumptions of the goals of the bureaucrats or entire organizations. There are models of political behavior dealing with strategic decision making both for international relations and domestic politics.

The foundation of all model building is the perhaps the arrogant statement that we know what caused the result we see. This is embodied in the simple mathematical statement of the functional form that $y = f(x)$. The response, Y, is caused by the stimulus, X. Every model will eventually come to this final place and it will be here that the theory will live or die. Will the data support this hypothesis? If so then fine, we shall believe this version of the world until a better theory comes to replace it. This is the process by which we moved from flat earth to round earth, from earth-center solar system to sun-center solar system, and on and on.

The scientific method does not confirm a theory for all time: it does not prove "truth". All theories are subject to review and may be overturned. These are lessons we learned as we first developed the concept of the hypothesis test earlier in this book. Here, as we begin this section, these concepts deserve review because the tool we will develop here is the cornerstone of the scientific method and the stakes are higher. Full theories will rise or fall because of this statistical tool: regression.

In this chapter we will begin with correlation, the investigation of relationships among variables that may or may not be founded on a cause and effect model. The variables simply move in the same, or opposite, direction. That is to say, they do not move randomly. Correlation provides a measure of the degree to which this is true. From there we develop a tool to measure cause and effect relationships: regression analysis. We will be able to formulate models and tests to determine if they are statistically sound. If they

are found to be so, then we can use them to make predictions: if as a matter of policy we changed the value of this variable what would happen to this other variable? If we imposed a gasoline tax of 50 cents per gallon how would that effect the carbon emissions, sales of Hummers/Hybrids, use of mass transit, etc.? The ability to provide answers to these types of questions is the value of regression as both a tool to help us understand our world and to make thoughtful policy decisions.

This page titled 10.1: Introduction is shared under a CC BY license and was authored, remixed, and/or curated by .

10.2: The Correlation Coefficient r

As we begin this section we note that the type of data we will be working with has changed. Perhaps unnoticed, all the data we have been using is for a single variable. It may be from two samples, but it is still a single variable (**univariate**) . The type of data described in the examples above and for any model of cause and effect is **bivariate** data ("bi" for two variables). In reality, statisticians use **multivariate** data, meaning they use many variables in their analyses.

For our work we can classify data into three broad categories, time series data, cross-section data, and panel data. We met the first two very early on. Time series data measures a single unit of observation; say a person, or a company or a country, as time passes. What are measured will be at least two characteristics, say the person's income, the quantity of a particular good they buy and the price they paid. This would be three pieces of information in one time period, say 1985. If we followed that person across time we would have those same pieces of information for 1985, 1986, 1987, etc. This would constitute a times series data set. If we did this for 10 years we would have 30 pieces of information concerning this person's consumption habits of this good for the past decade and we would know their income and the price they paid.

A second type of data set is for cross-section data. Here the variation is not across time for a single unit of observation, but across units of observation during one point in time. For a particular period of time we would gather the price paid, amount purchased, and income of many individual people.

A third type of data set is panel data. Here a panel of units of observation is followed across time. If we take our example from above we might follow 500 people, the unit of observation, through time, ten years, and observe their income, price paid and quantity of the good purchased. If we had 500 people and data for ten years for price, income and quantity purchased we would have 15,000 pieces of information. These types of data sets are very expensive to construct and maintain. They do, however, provide a tremendous amount of information that can be used to answer very important questions. As an example, what is the effect on the labor force participation rate of women as their family of origin, mother and father, age? Or are there differential effects on health outcomes depending upon the age at which a person started smoking? Only panel data can give answers to these and related questions because we must follow multiple people across time. The work we do here however will not be fully appropriate for data sets such as these.

Beginning with a set of data with two independent variables we ask the question: are these related? One way to visually answer this question is to create a scatter plot of the data. We could not do that before when we were doing descriptive statistics because those data were univariate. Now we have bivariate data so we can plot in two dimensions.

To provide mathematical precision to the measurement of what we see we use the correlation coefficient. The correlation tells us something about the co-movement of two variables, but nothing about why this movement occurred. Formally, correlation analysis assumes that both variables being analyzed are **independent** variables. This means that neither one causes the movement in the other. Further, it means that neither variable is dependent on the other, or for that matter, on any other variable. Even with these limitations, correlation analysis can yield some interesting results.

The correlation coefficient, ρ (pronounced rho), is the mathematical statistic for a population that provides us with a measurement of the strength of a linear relationship between the two variables. For a sample of data, the statistic, r, developed by Karl Pearson in the early 1900s, is an estimate of the population correlation and is defined mathematically as:

$$r_{XY} = \frac{\sum(X_i - \overline{X}) * (Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 * \sum (Y_i - \overline{Y})^2}}$$

or

$$r_{XY} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sqrt{\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right] * \left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right]}}$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of the two independent variables X and Y, X_i and Y_i are the individual observations of X and Y, and n is our sample size. The correlation coefficient r ranges in value from -1 to 1. The second equivalent formula is often used because it may be computationally easier. As scary as these formulas look, they are really just the ratio of the covariance between the two variables and the product of their two standard deviations. That is to say, it is a measure of relative variances.

To visualize any **linear** relationship that may exist review the plot of a scatter diagrams of the standardized data. Figure 10.2.1 presents several scatter diagrams and the calculated value of r. In panels (a) and (b) notice that the data generally trend together, (a) upward and (b) downward. Panel (a) is an example of a positive correlation and panel (b) is an example of a negative correlation, or relationship. The sign of the correlation coefficient tells us if the relationship is a positive or negative (inverse) one. If all the values of two variables X_1 and X_2 are on a straight line the correlation coefficient will be either 1 or -1 depending on whether the line has a positive or negative slope and the closer to one or negative one the stronger the relationship between the two variables.

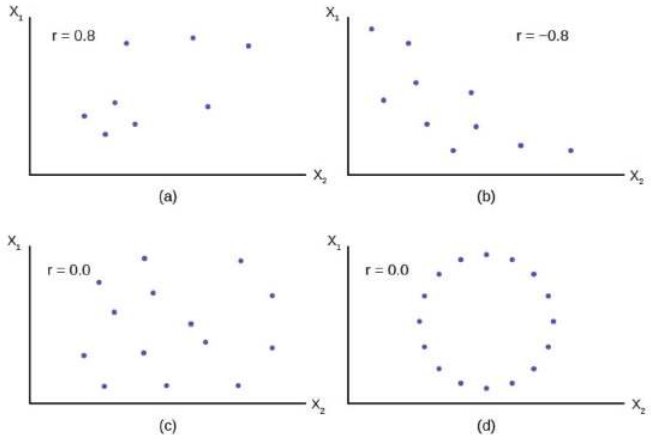


Figure 10.2.1

Remember, all the correlation coefficient tells us is whether or not the data are linearly related. In panel (d) the variables obviously have some type of very specific relationship to each other, but the correlation coefficient is zero, indicating no **linear** relationship exists. In panel (c) the variables are not related and the correlation coefficient is understandably equal to zero.

If you suspect a linear relationship between two variables X and Y, then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

- The value of r is always between –1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the **linear** relationship between variables X and Y. Values of r close to –1 or to $+1$ indicate a stronger linear relationship between the two variables.
- If $r = 0$ there is absolutely no linear relationship between variables X and Y **(no linear correlation)**.
- 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)**.

Note

Strong correlation does not suggest that X causes Y or that Y causes X. We say **"correlation does not imply causation."**

This page titled 10.2: The Correlation Coefficient r is shared under a CC BY license and was authored, remixed, and/or curated by .

10.3: Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between X and Y.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- ρ = population correlation coefficient (unknown)
- r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between variables X and Y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between variables X and Y. If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

Performing the Hypothesis Test

- **Null Hypothesis:** $H_0 : \rho = 0$
- **Alternate Hypothesis:** $H_a : \rho \neq 0$

What the Hypotheses Mean in Words:

- **Null Hypothesis** H_0: The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between X and Y in the population.
- **Alternate Hypothesis** H_a: The population correlation coefficient is significantly different from zero. There is a significant linear relationship (correlation) between X and Y in the population.

Drawing a Conclusion There are two methods of making the decision concerning the hypothesis. The test statistic to test this hypothesis is:

$$t_{obs} = \frac{r - \rho}{\sqrt{(1 - r^2)/(n - 2)}}$$

OR

$$t_{obs} = \frac{(r - \rho) * \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

Where the second formula is an equivalent form of the test statistic, n is the sample size and the degrees of freedom are $n - 2$. This is a t-statistic and operates in the same way as other t-tests. Calculate the t-value and compare that with the critical value from the t -table at the appropriate degrees of freedom and the level of confidence you wish to maintain. If the calculated (observed) value is in the tail, then reject the null hypothesis that there is no linear relationship between these two independent random variables. If the calculated (observed) t-value is NOT in the tail, then we cannot reject the null hypothesis that there is no linear relationship between the two variables.

A quick shorthand way to test correlations is the relationship between the sample size and the correlation. If:

$$|r| \geq \frac{2}{\sqrt{n}}$$

then this implies that the correlation between the two variables demonstrates that a linear relationship exists and is statistically significant at approximately the 0.05 level of significance. As the formula indicates, there is an inverse relationship between the

sample size and the required correlation for significance of a linear relationship. With only 10 observations, the required correlation for significance is 0.6325, for 30 observations the required correlation for significance decreases to 0.3651 and at 100 observations the required level is only 0.2000.

Correlations may be helpful in visualizing the data, but are not appropriately used to "explain" a relationship between two variables. Perhaps no single statistic is more misused than the correlation coefficient. Citing correlations between health conditions and everything from place of residence to eye color have the effect of implying a cause and effect relationship. This simply cannot be accomplished with a correlation coefficient. The correlation coefficient is, of course, innocent of this misinterpretation. It is the duty of the analyst to use a statistic that is designed to test for cause and effect relationships and report only those results if they are intending to make such a claim. The problem is that passing this more rigorous test is difficult so lazy and/or unscrupulous "researchers" fall back on correlations when they cannot make their case legitimately.

This page titled 10.3: Testing the Significance of the Correlation Coefficient is shared under a CC BY license and was authored, remixed, and/or curated by .

10.4: Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$Y = a + bX$$

where a and b are constant numbers.

The variable X **is the independent variable, and** Y **is the dependent variable.** Another way to think about this equation is a statement of cause and effect. The X variable is the cause and the Y variable is the hypothesized effect. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 10.4.1

The following examples are linear equations.

$Y = 3 + 2X$

$Y = -0.01 + 1.2X$

The graph of a linear equation of the form $Y = a + bX$ is a **straight line**. Any line that is not vertical can be described by this equation

Example 10.4.2

Graph the equation $Y = -1 + 2X$.

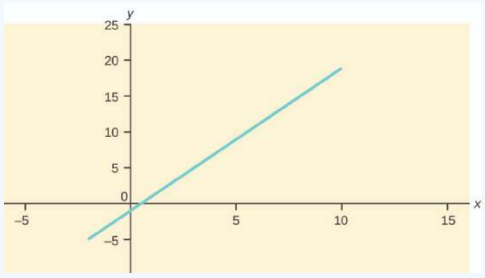


Figure 10.4.1

Exercise 10.4.1

Is the following an example of a linear equation? Why or why not?

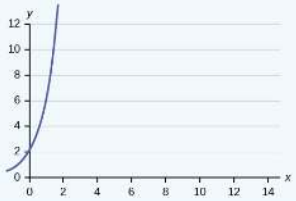


Figure 10.4.2

Example 10.4.3

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Answer

Let X = the number of hours it takes to get the job done.
Let Y = the total cost to the customer.

The \$31.50 is a fixed cost. If it takes X hours to complete the job, then (32)(X) is the cost of the word processing only. The total cost is: $Y = 31.50 + 32X$

Slope and Y-Intercept of a Linear Equation

For the linear equation $Y = a + bX$, b = slope and $a = Y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the Y-intercept is the Y coordinate of the point $(0, a)$ where the line crosses the y-axis. From calculus the slope is the first derivative of the function. For a linear function, the slope is $dY/dX = b$ where we can read the mathematical expression as "the change in $Y(dY)$ that results from a unit change in $X(dX)$ equals b".



Figure 10.4.3 Three possible graphs of $Y = a + bX$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 10.4.4

Sven tutors to make extra money for college. For each tutoring session, he charges a one-time fee of \$25 plus \$15 per hour of tutoring. A linear equation that expresses the total amount of money Sven earns for each session he tutors is $Y = 25 + 15X$.

What are the independent and dependent variables? What is the Y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The independent variable (X) is the number of hours Sven tutors each session. The dependent variable (Y) is the amount, in dollars, Sven earns for each session.

The Y-intercept is 25 $(a = 25)$. At the start of the tutoring session, Sven charges a one-time fee of \$25 (this is when $X = 0$). The slope is 15 $(b = 15)$. For each session, Sven earns \$15 for each hour he tutors.

This page titled 10.4: Linear Equations is shared under a CC BY license and was authored, remixed, and/or curated by .

10.5: The Regression Equation

Regression analysis is a statistical technique that can test the hypothesis that a variable is dependent upon one or more other variables. Further, regression analysis can provide an estimate of the magnitude of the impact of a change in one variable on another. This last feature, of course, is all important in predicting future values.

Regression analysis is based upon a functional relationship among variables and further, assumes that the relationship is linear. This linearity assumption is required because, for the most part, the theoretical statistical properties of non-linear estimation are not well worked out yet by the mathematicians and econometricians. This presents us with some difficulties in economic analysis because many of our theoretical models are nonlinear. The marginal cost curve, for example, is decidedly nonlinear as is the total cost function, if we are to believe in the effect of specialization of labor and the law of diminishing marginal product. There are techniques for overcoming some of these difficulties, exponential and logarithmic transformation of the data for example, but at the outset we must recognize that standard ordinary least squares (OLS) regression analysis will always use a linear function to estimate what might be a nonlinear relationship.

The general linear regression model can be stated by the equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

where β_0 is the intercept, β_i's are the slope between Y and the appropriate X_i, and ϵ (pronounced epsilon) is the error term that captures errors in measurement of Y and the effect on Y of any variables missing from the equation that would contribute to explaining variations in Y. This equation is the theoretical population equation and therefore uses Greek letters. The equation we will estimate will have the Roman equivalent symbols. This is parallel to how we kept track of the population parameters and sample parameters before. The symbol for the population mean was μ and for the sample mean $\overline{x}$ and for the population standard deviation was σ and for the sample standard deviation was s. The equation that will be estimated with a sample of data for two independent variables will thus be:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

As with our earlier work with probability distributions, this model works only if certain assumptions hold. These are that the Y is normally distributed, the errors are also normally distributed with a mean of zero and a constant standard deviation, and that the error terms are independent of the size of X and independent of each other.

Assumptions of the Ordinary Least Squares Regression Model

Each of these assumptions needs a bit more explanation. If one of these assumptions fails to be true, then it will have an effect on the quality of the estimates. Some of the failures of these assumptions can be fixed while others result in estimates that quite simply provide no insight into the questions the model is trying to answer or worse, give biased estimates.

1. The independent variables, X_i, are all measured without error, and are fixed numbers that are independent of the error term. This assumption is saying in effect that Y is deterministic, the result of a fixed component "X" and a random error component "ϵ."
2. The error term is a random variable with a mean of zero and a constant variance. The meaning of this is that the variances of the independent variables are independent of the value of the variable. Consider the relationship between personal income and the quantity of a good purchased as an example of a case where the variance is dependent upon the value of the independent variable, income. It is plausible that as income increases the variation around the amount purchased will also increase simply because of the flexibility provided with higher levels of income. The assumption is for constant variance with respect to the magnitude of the independent variable called homoscedasticity. If the assumption fails, then it is called heteroscedasticity. Figure 10.5.1 shows the case of homoscedasticity where all three distributions have the same variance around the predicted value of Y regardless of the magnitude of X.
3. While the independent variables are all fixed values they are from a probability distribution that is normally distributed. This can be seen in Figure 10.5.1 by the shape of the distributions placed on the predicted line at the expected value of the relevant value of Y.
4. The independent variables are independent of Y, but are also assumed to be independent of the other X variables. The model is designed to estimate the effects of independent variables on some dependent variable in accordance with a proposed theory. The

case where some or more of the independent variables are correlated is not unusual. There may be no cause and effect relationship among the independent variables, but nevertheless they move together. Take the case of a simple supply curve where quantity supplied is theoretically related to the price of the product and the prices of inputs. There may be multiple inputs that may over time move together from general inflationary pressure. The input prices will therefore violate this assumption of regression analysis. This condition is called multicollinearity, which will be taken up in detail later.

5. The error terms are uncorrelated with each other. This situation arises from an effect on one error term from another error term. While not exclusively a time series problem, it is here that we most often see this case. An X variable in time period one has an effect on the Y variable, but this effect then has an effect in the next time period. This effect gives rise to a relationship among the error terms. This case is called autocorrelation, “self-correlated.” The error terms are now not independent of each other, but rather have their own effect on subsequent error terms.

Figure 10.5.1 shows the case where the assumptions of the regression model are being satisfied. The estimated line is $\hat{Y} = a + bX$. Three values of X are shown. A normal distribution is placed at each point where X equals the estimated line and the associated error at each value of Y. Notice that the three distributions are normally distributed around the point on the line, and further, the variation, variance, around the predicted value is constant indicating homoscedasticity from assumption 2. Figure 10.5.1 does not show all the assumptions of the regression model, but it helps visualize these important ones.

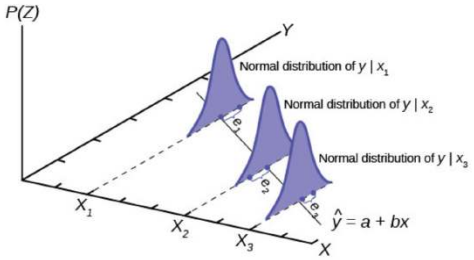


Figure 10.5.1

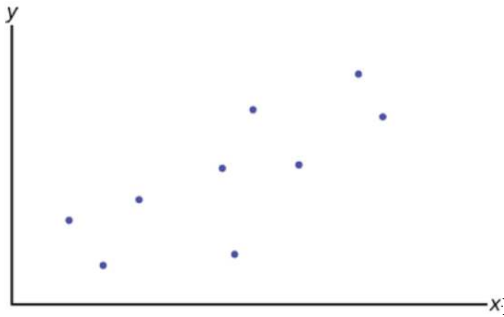


Figure 10.5.2

This is the general form that is most often called the multiple regression model. So-called "simple" regression analysis has only one independent (right-hand) variable rather than many independent variables. Simple regression is just a special case of multiple regression. There is some value in beginning with simple regression: it is easy to graph in two dimensions, difficult to graph in three dimensions, and impossible to graph in more than three dimensions. Consequently, our graphs will be for the simple regression case. Figure 10.5.2 presents the regression problem in the form of a scatter plot graph of the data set where it is hypothesized that Y is dependent upon the single independent variable X.

A basic relationship from principles of microeconomics is the consumption function. This theoretical relationship states that as a person's income rises, their consumption rises, but by a smaller amount than the rise in income. If Y is consumption and X is income in the equation below Figure 10.5.2, the regression problem is, first, to establish that this relationship exists, and second, to determine the impact of a change in income on a person's consumption. The parameter β_1 is called the marginal propensity to consume (MPC) in economics.

Each "dot" in Figure 10.5.2 represents the consumption and income of different individuals at some point in time. This was called cross-section data earlier; observations on variables at one point in time across different people or other units of measurement. This analysis is often done with time series data, which would be the consumption and income of one individual or country at different

points in time. For macroeconomic problems it is common to use times series aggregated data for a whole country. For this particular theoretical concept these data are readily available in the annual report of the President's Council of Economic Advisors.

The regression problem comes down to determining which straight line would best represent the data in Figure 10.5.3. Regression analysis is sometimes called "least squares" analysis because the method of determining which line best "fits" the data is to minimize the sum of the squared residuals of a line put through the data.

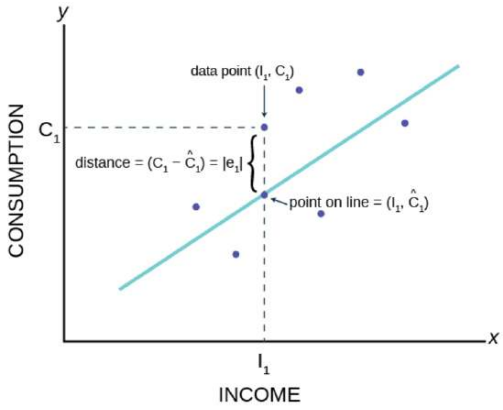


Figure 10.5.3
Population Equation: $\mathrm{C} = \beta_0 + \beta_1 \mathrm{Income} + \varepsilon$
Estimated Equation: $C = b_0 + b_1 \mathrm{Income} + e$

This figure shows the assumed relationship between consumption and income from microeconomic theory. Here the data are plotted as a scatter plot and an estimated straight line has been drawn. From this graph we can see an error term, e_1. Each data point also has an error term. Again, the error term is put into the equation to capture effects on consumption that are not caused by income changes. Such other effects might be a person's savings or wealth, or periods of unemployment. We will see how by minimizing the sum of these errors we can get an estimate for the slope and intercept of this line.

Consider the graph below. The notation has returned to that for the more general model rather than the specific case of the consumption function in our example.

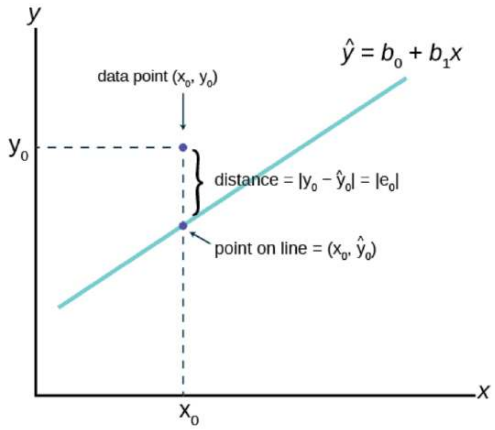


Figure 10.5.4

The $\hat{Y}$ is read **"Y hat"** and is the **estimated value of** Y. (In Figure 10.5.3 $\hat{C}$ represents the estimated value of consumption because it is on the estimated line.) It is the value of Y obtained using the regression line. $\hat{Y}$ is not generally equal to Y from the data.

The term $Y_0 - \hat{Y}_0 = e_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The error term was put into the estimating equation to capture missing variables and errors in measurement that may have occurred in the dependent variables. The **absolute value of a residual** measures the vertical distance between the actual and the estimated value of Y. In other words, it measures the vertical distance between the actual data point Y_0 and the predicted point $\hat{Y}$ on the line as can be seen on the graph at point X_0.

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 graph, $Y_0 - \hat{Y}_0 = e_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive. For each data point the residuals, or errors, are calculated $Y_i - \hat{Y}_i = e_i$ for $i = 1, 2, 3, ..., n$, where n is the sample size. Each $|e|$ is a vertical distance.

The sum of the errors squared is the term called **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the straight line that has the parameter values of b_0 and b_1 that minimizes the **SSE**. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{Y} = b_0 + b_1 X$$

where:

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$b_1 = \frac{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma (X_i - \overline{X})^2}$$

or

$$b_1 = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

The sample means of the X values and the Y values are $\overline{X}$ and $\overline{Y}$, respectively. The best fit line always passes through the point ($\overline{Y}$, $\overline{X}$) called the points of means.

The slope b_1 can also be written as:

$$b_1 = r_{XY} \left(\frac{s_Y}{s_X} \right)$$

where s_Y = the standard deviation of the Y values and s_X = the standard deviation of the X values and r is the correlation coefficient between variables X and Y.

These equations are called the Normal Equations and come from another very important mathematical finding called the Gauss-Markov Theorem without which we could not do regression analysis. The Gauss-Markov Theorem tells us that the estimates we get from using the ordinary least squares (OLS) regression method will result in estimates that have some very important properties. In the Gauss-Markov Theorem it was proved that a least squares line is BLUE, which is, **B**est, **L**inear, **U**nbiased, **E**stimator. Best is the statistical property that an estimator is the one with the minimum variance. Linear refers to the property of the type of line being estimated. An unbiased estimator is one whose estimating function has an expected mean equal to the mean of the population. (You will remember that the expected value of $\mu_{\overline{x}}$ was equal to the population mean μ in accordance with the Central Limit Theorem. This is exactly the same concept here).

Both Gauss and Markov were giants in the field of mathematics, and Gauss in physics too, in the 18th century and early 19th century. They barely overlapped chronologically and never in geography, but Markov's work on this theorem was based extensively on the earlier work of Carl Gauss. The extensive applied value of this theorem had to wait until the middle of this last century.

Using the OLS method we can now find the **estimate of the error variance** which is the variance of the squared errors, e^2. This is sometimes called the **standard error of the estimate**. (Grammatically this is probably best said as the estimate of the **error's** variance) The formula for the estimate of the error variance is:

$$s_e^2 = \frac{\Sigma \left(Y_i - \hat{Y}_i \right)^2}{n - k} = \frac{\Sigma e_i^2}{n - k}$$

where $\hat{Y}$ is the predicted value of Y and Y is the observed value, and thus the term $\left(Y_i - \hat{Y}_i\right)^2$ is the squared errors that are to be minimized to find the estimates of the regression line parameters. This is really just the variance of the error terms and follows our regular variance formula. One important note is that here we are dividing by $(n-k)$, which is the degrees of freedom. The degrees of freedom of a regression equation will be the number of observations, n, reduced by the number of estimated parameters, k, which includes the intercept as a parameter.

The variance of the errors is fundamental in testing hypotheses for a regression. It tells us just how "tight" the dispersion is about the line. As we will see shortly, the greater the dispersion about the line, meaning the larger the variance of the errors, the less probable that the hypothesized independent variable will be found to have a significant effect on the dependent variable. In short, the theory being tested will more likely fail if the variance of the error term is high. Upon reflection this should not be a surprise. As we tested hypotheses about a mean we observed that large variances reduced the calculated test statistic and thus it failed to reach the tail of the distribution. In those cases, the null hypotheses could not be rejected. If we cannot reject the null hypothesis in a regression problem, we must conclude that the hypothesized independent variable has no effect on the dependent variable.

A way to visualize this concept is to draw two scatter plots of X and Y data along a predetermined line. The first will have little variance of the errors, meaning that all the data points will move close to the line. Now do the same except the data points will have a large estimate of the error variance, meaning that the data points are scattered widely along the line. Clearly the confidence about a relationship between X and Y is effected by this difference between the estimate of the error variance.

Testing the Parameters of the Line

The whole goal of the regression analysis was to test the hypothesis that the dependent variable, Y, was in fact dependent upon the values of the independent variables as asserted by some foundation theory, such as the consumption function example. Looking at the estimated equation under Figure 10.5.3, we see that this amounts to determining the values of b_0 and b_1. Notice that again we are using the convention of Greek letters for the population parameters and Roman letters for their estimates.

The regression analysis output provided by the computer software will produce an estimate of b_0 and b_1, and any other b's for other independent variables that were included in the estimated equation. The issue is how good are these estimates? In order to test a hypothesis concerning any estimate, we have found that we need to know the underlying sampling distribution. It should come as no surprise at his stage in the course that the answer is going to be the normal distribution. This can be seen by remembering the assumption that the error term in the population, ϵ, is normally distributed. If the error term is normally distributed and the variance of the estimates of the equation parameters, b_0 and b_1, are determined by the variance of the error term, it follows that the variances of the parameter estimates are also normally distributed. And indeed this is just the case.

We can see this by the creation of the test statistic for the test of hypothesis for the slope parameter, β_1 in our consumption function equation. To test whether or not Y does indeed depend upon X, or in our example, that consumption depends upon income, we need only test the hypothesis that β_1 equals zero. This hypothesis would be stated formally as:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

If we cannot reject the null hypothesis, we must conclude that our theory has no validity. If we cannot reject the null hypothesis that $\beta_1 = 0$ then b_1, the coefficient of Income, is zero and zero times anything is zero. Therefore the effect of Income on Consumption is zero. There is no relationship as our theory had suggested.

Notice that we have set up the presumption, the null hypothesis, as "no relationship". This puts the burden of proof on the alternative hypothesis. In other words, if we are to validate our claim of finding a relationship, we must do so with a level of significance greater than 90, 95, or 99 percent. The status quo is ignorance, no relationship exists, and to be able to make the claim that we have actually added to our body of knowledge we must do so with significant probability of being correct.

The test statistic for this test comes directly from our old friend the standardizing formula:

$$t_{obs} = \frac{b_1 - \beta_1}{s_{b_1}}$$

where b_1 is the estimated value of the slope of the regression line, β_1 is the hypothesized value of beta, in this case zero, and s_{b_1} is the standard deviation of the estimate of b_1. In this case we are asking how many standard deviations is the estimated slope away from the hypothesized slope. This is exactly the same question we asked before with respect to a hypothesis about a mean: how many standard deviations is the estimated mean, the sample mean, from the hypothesized mean?

The test statistic is written as a Student's t-distribution, but if the sample size is larger enough so that the degrees of freedom are greater than 100 we may again use the normal distribution. To see why we can use the Student's t or normal distribution we have only to look at s_{b_1},the formula for the standard deviation of the estimate of b_1:

$$s_{b_1} = \frac{s_e^2}{\sqrt{\left(X_i - \overline{X}\right)^2}}$$

or

$$s_{b_1} = \frac{s_e^2}{(n-1)s_X^2}$$

Where s_e is the estimate of the error variance and s_X^2 is the variance of X values of the coefficient of the independent variable being tested.

We see that s_e, the **estimate of the error variance**, is part of the computation. Because the estimate of the error variance is based on the assumption of normality of the error terms, we can conclude that the sampling distribution of the b's, the coefficients of our hypothesized regression line, are also normally distributed.

One last note concerns the degrees of freedom of the test statistic, $df = n - k$. Previously we subtracted 1 from the sample size to determine the degrees of freedom in a Student's t problem. Here we must subtract one degree of freedom for each parameter estimated in the equation. For the example of the consumption function we lose 2 degrees of freedom, one for b_0, the intercept, and one for b_1, the slope of the consumption function. The degrees of freedom would be $n - k - 1$, where k is the number of independent variables and the extra one is lost because of the intercept. If we were estimating an equation with three independent variables, we would lose 4 degrees of freedom: three for the independent variables, k, and one more for the intercept.

The decision rule for the rejection of the null hypothesis follows exactly the same form as in all our previous test of hypothesis. Namely, if the calculated value of t (or z) falls into the tails of the distribution, where the tails are defined by α, the required significance level in the test, we reject the null hypothesis. If on the other hand, the calculated value of the test statistic is within the critical region, we cannot reject the null hypothesis.

If we conclude that we reject the null hypothesis, we are able to state with $(1 - \alpha)$ level of confidence that the slope of the line is given by b_1. This is an extremely important conclusion. Regression analysis not only allows us to test if a relationship exists, but we can also determine the magnitude of that relationship, if one is found to exist. It is this feature of regression analysis that makes it so valuable. If models can be developed that have statistical validity, we are then able to simulate the effects of changes in variables that may be under our control with some degree of probability, of course. For example, if advertising is demonstrated to effect sales, we can determine the effects of changing the advertising budget and decide if the increased sales are worth the added expense.

Multicollinearity

Our discussion earlier indicated that like all statistical models, the OLS regression model has important assumptions attached. Each assumption, if violated, has an effect on the ability of the model to provide useful and meaningful estimates. The Gauss-Markov Theorem has assured us that the OLS estimates are unbiased and minimum variance, but this is true only under the assumptions of the model. Here we will look at the effects on OLS estimates if the independent variables are correlated. The other assumptions and the methods to mitigate the difficulties they pose if they are found to be violated are examined in econometrics courses. We take up multicollinearity because it is so often prevalent in economic models and it often leads to frustrating results.

The OLS model assumes that all the independent variables are independent of each other. This assumption is easy to test for a particular sample of data with simple correlation coefficients. Correlation, like much in statistics, is a matter of degree: a little is not good, and a lot is terrible.

The goal of the regression technique is to tease out the independent impacts of each of a set of independent variables on some hypothesized dependent variable. If two 2 independent variables are interrelated, that is, correlated, then we cannot isolate the effects on Y of one from the other. In an extreme case where X_1 is a linear combination of X_2, correlation equal to one, both variables move in identical ways with Y. In this case it is impossible to determine the variable that is the true cause of the effect on Y. (If the two variables were actually perfectly correlated, then mathematically no regression results could actually be calculated.)

The normal equations for the coefficients show the effects of multicollinearity on the coefficients.

$$b_1 = \frac{s_Y \left(r_{X_1 Y} - r_{X_1 X_2} r_{X_2 Y}\right)}{s_{X_1} \left(1 - r^2_{X_1 X_2}\right)}$$

$$b_2 = \frac{s_Y \left(r_{X_2 Y} - r_{X_1 X_2} r_{X_1 Y}\right)}{s_{X_2} \left(1 - r^2_{X_1 X_2}\right)}$$

$$b_0 = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

The correlation between X_1 and X_2, $r^2_{X_1 X_2}$, appears in the denominator of both the estimating formula for b_1 and b_2. If the assumption of independence holds, then this term is zero. This indicates that there is no effect of the correlation on the coefficient. On the other hand, as the correlation between the two independent variables increases the denominator decreases, and thus the estimate of the coefficient increases. The correlation has the same effect on both of the coefficients of these two variables. In essence, each variable is “taking” part of the effect on Y that should be attributed to the collinear variable. This results in biased estimates.

Multicollinearity has a further deleterious impact on the OLS estimates. The correlation between the two independent variables also shows up in the formulas for the estimate of the variance for the coefficients.

$$s^2_{b_1} = \frac{s^2_e}{(n-1) s^2_{X_1} \left(1 - r^2_{X_1 X_2}\right)}$$

$$s^2_{b_2} = \frac{s^2_e}{(n-1) s^2_{X_2} \left(1 - r^2_{X_1 X_2}\right)}$$

Here again we see the correlation between X_1 and X_2 in the denominator of the estimates of the variance for the coefficients for both variables. If the correlation is zero as assumed in the regression model, then the formula collapses to the familiar ratio of the variance of the errors to the variance of the relevant independent variable. If however the two independent variables are correlated, then the variance of the estimate of the coefficient increases. This results in a smaller t-value for the test of hypothesis of the coefficient. In short, multicollinearity results in failing to reject the null hypothesis that the X variable has no impact on Y when in fact X does have a statistically significant impact on Y. Said another way, the large standard errors of the estimated coefficient created by multicollinearity suggest statistical insignificance even when the hypothesized relationship is strong.

How Good is the Equation?

In the last section we concerned ourselves with testing the hypothesis that the dependent variable did indeed depend upon the hypothesized independent variable or variables. It may be that we find an independent variable that has some effect on the dependent variable, but it may not be the only one, and it may not even be the most important one. Remember that the error term was placed in the model to capture the effects of any missing independent variables. It follows that the error term may be used to give a measure of the "goodness of fit" of the equation taken as a whole in explaining the variation of the dependent variable, Y.

The **multiple correlation coefficient**, also called the **coefficient of multiple determination** or the **coefficient of determination**, is given by the formula:

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

where SSR is the regression sum of squares, the squared deviation of the predicted value of Y from the mean value of $Y (\hat{Y} - \overline{Y})$, and SST is the total sum of squares which is the total squared deviation of the dependent variable, Y, from its mean value,

including the error term, SSE, the sum of squared errors. Figure 10.5.5 shows how the total deviation of the dependent variable, Y, is partitioned into these two pieces.

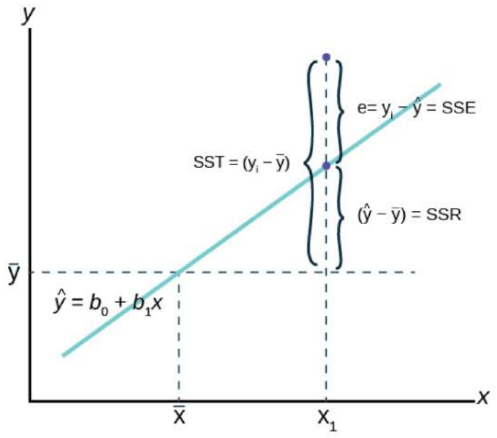


Figure 10.5.5

Figure 10.5.5 shows the estimated regression line and a single observation, X_1. Regression analysis tries to explain the variation of the data about the mean value of the dependent variable, Y. The question is, why do the observations Y vary from the average level of Y? The value of Y at observation X_1 varies from the mean of Y by the difference $\left(Y_i - \overline{Y}\right)$. The sum of these differences squared is SST, the sum of squares total. The actual value of Y at X_1 deviates from the estimated value, $\hat{Y}$, by the difference between the estimated value and the actual value, $\left(Y_i - \hat{Y}\right)$. We recall that this is the error term, e, and the sum of these errors is SSE, sum of squared errors. The deviation of the predicted value of Y, $\hat{Y}$, from the mean value of Y is $(\hat{Y} - \overline{Y})$ and is the SSR, sum of squares regression. It is called "regression" because it is the deviation explained by the regression. (Sometimes the SSR is called SSM for sum of squares mean because it measures the deviation from the mean value of the dependent variable, Y, as shown on the graph.).

Because the SST = SSR + SSE we see that the multiple correlation coefficient is the percent of the variance, or deviation in Y from its mean value, that is explained by the equation when taken as a whole. R^2 will vary between zero and 1, with zero indicating that none of the variation in Y was explained by the equation and a value of 1 indicating that 100% of the variation in Y was explained by the equation. For time series studies expect a high R^2 and for cross-section data expect low R^2.

While a high R^2 is desirable, remember that it is the tests of the hypothesis concerning the existence of a relationship between a set of independent variables and a particular dependent variable that was the motivating factor in using the regression model. It is validating a cause and effect relationship developed by some theory that is the true reason that we chose the regression analysis. Increasing the number of independent variables will have the effect of increasing R^2. To account for this effect the proper measure of the coefficient of determination is the $\overline{R}^2$, adjusted for degrees of freedom, to keep down mindless addition of independent variables.

There is no statistical test for the R^2 and thus little can be said about the model using R^2 with our characteristic confidence level. Two models that have the same size of SSE, that is sum of squared errors, may have very different R^2 if the competing models have different SST, total sum of squared deviations. The goodness of fit of the two models is the same; they both have the same sum of squares unexplained, errors squared, but because of the larger total sum of squares on one of the models the R^2 differs. Again, the real value of regression as a tool is to examine hypotheses developed from a model that predicts certain relationships among the variables. These are tests of hypotheses on the coefficients of the model and not a game of maximizing R^2.

Another way to test the general quality of the overall model is to test the coefficients as a group rather than independently. Because this is multiple regression (more than one X), we use the F-test to determine if our coefficients collectively affect Y. The hypothesis is:

$H_o : \beta_1 = \beta_2 = ... = \beta_i = 0$

H_a: "at least one of the β_i is not equal to 0"

If the null hypothesis cannot be rejected, then we conclude that none of the independent variables contribute to explaining the variation in Y. Reviewing Figure 10.5.5 we see that SSR, the explained sum of squares, is a measure of just how much of the variation in Y is explained by all the variables in the model. SSE, the sum of the errors squared, measures just how much is unexplained. It follows that the ratio of these two can provide us with a statistical test of the model as a whole. Remembering that the F-distribution is a ratio of chi-squared distributions and that variances are distributed according to chi-squared, and the sum of squared errors and the sum of squares are both variances, we have the test statistic for this hypothesis as:

$$F_{obs} = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)}$$

where n is the number of observations and k is the number of independent variables. It can be shown that this is equivalent to:

$$F_{obs} = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}$$

Figure 10.5.5 where R^2 is the coefficient of determination which is also a measure of the "goodness" of the model.

As with all our tests of hypothesis, we reach a conclusion by comparing the calculated F-statistic with the critical value given our desired level of confidence. If the calculated test statistic, an F-statistic in this case, is in the tail of the distribution, then we need to reject the null hypothesis. By rejecting the null hypothesis, we conclude that this specification of this model has validity, because at least one of the estimated coefficients is significantly different from zero.

An alternative way to reach this conclusion is to use the p-value comparison rule. The p-value is the area in the tail, given the calculated F-statistic. In essence, the computer is finding the F-value in the table for us. The computer regression output for the observed F-statistic is typically found in the ANOVA table section labeled "significance F". How to read the output of an Excel regression is presented below. This is the probability of rejecting a false null hypothesis. If this probability is less than our pre-determined alpha error, then the conclusion is that we reject the null hypothesis.

Dummy Variables

Thus far the analysis of the OLS regression technique assumed that the independent variables in the models tested were continuous random variables. There are, however, no restrictions in the regression model against independent variables that are binary. This opens the regression model for testing hypotheses concerning categorical variables such as gender, race, region of the country, before a certain data, after a certain date and innumerable others. These categorical variables take on only two values, 1 and 0, success or failure, from the binomial probability distribution. The form of the equation becomes:

$$\hat{Y} = b_0 + b_2 X_2 + b_1 X_1$$

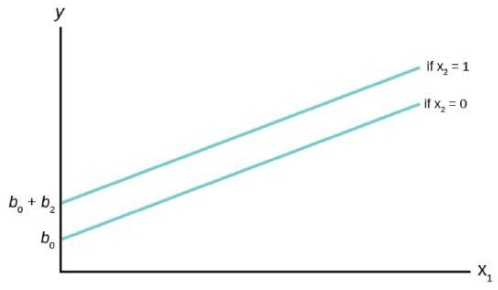


Figure 10.5.6

where X_2 is the dummy variable and X_1 is some continuous random variable. The constant, b_0, is the Y-intercept, the value where the line crosses the y-axis. When the value of $X_2 = 0$, the estimated line crosses at b_0. When the value of $X_2 = 1$ then the estimated line crosses at $b_0 + b_2$. In effect the dummy variable causes the estimated line to shift either up or down by the size of the effect of the characteristic captured by the dummy variable. Note that this is a simple parallel shift and does not affect the impact of the other independent variable, X_1.This variable is a continuous random variable and predicts different values of Y at different values of X_1 holding constant the condition of the dummy variable.

An example of the use of a dummy variable is the work estimating the impact of gender on salaries. There is a full body of literature on this topic and dummy variables are used extensively. For this example the salaries of elementary and secondary school teachers for a particular state is examined. Using a homogeneous job category, school teachers, and for a single state reduces many of the variations that naturally effect salaries such as differential physical risk, cost of living in a particular state, and other working conditions. The estimating equation in its simplest form specifies salary as a function of various teacher characteristic that economic theory would suggest could affect salary. These would include education level as a measure of potential productivity, age and/or experience to capture on-the-job training, again as a measure of productivity. Because the data are for school teachers employed in a public school districts rather than workers in a for-profit company, the school district's average revenue per average daily student attendance is included as a measure of ability to pay. The results of the regression analysis using data on 24,916 school teachers are presented below.

Variable	Regression Coefficients (b)	Standard Errors of the estimates for teacher's earnings function (s_b)
Intercept	4269.9	
Gender (male = 1)	632.38	13.39
Total Years of Experience	52.32	1.10
Years of Experience in Current District	29.97	1.52
Education	629.33	13.16
Total Revenue per ADA	90.24	3.76
R^2	.725	
n	24,916	

Table 10.5.1 Earnings Estimate for Elementary and Secondary School Teachers

The coefficients for all the independent variables are significantly different from zero as indicated by the standard errors. Dividing the standard errors of each coefficient results in a t-value greater than 1.96 which is the required level for 95% significance. The binary variable, our dummy variable of interest in this analysis, is gender where male is given a value of 1 and female given a value of 0. The coefficient is significantly different from zero with a dramatic t-statistic of 47 standard deviations. We thus reject the null hypothesis that the coefficient is equal to zero. Therefore we conclude that there is a premium paid male teachers of $632 after holding constant experience, education and the wealth of the school district in which the teacher is employed. It is important to note that these data are from some time ago and the $632 represents a six percent salary premium at that time. A graph of this example of dummy variables is presented below.

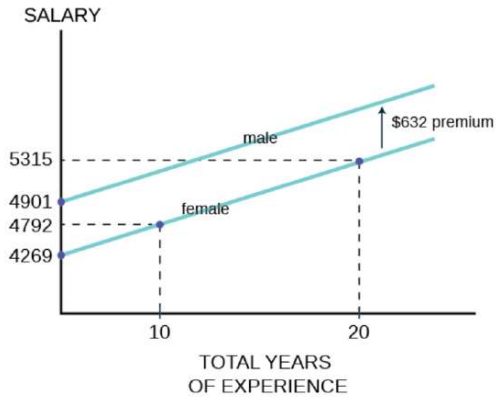


Figure 10.5.7

In two dimensions, salary is the dependent variable on the vertical axis and total years of experience was chosen for the continuous independent variable on horizontal axis. Any of the other independent variables could have been chosen to illustrate the effect of the dummy variable. The relationship between total years of experience has a slope of $52.32 per year of experience and the estimated line has an intercept of $4,269 if the gender variable is equal to zero, for female. If the gender variable is equal to 1, for male, the coefficient for the gender variable is added to the intercept and thus the relationship between total years of experience and salary is shifted upward parallel as indicated on the graph. Also marked on the graph are various points for reference. A female school teacher with 10 years of experience receives a salary of $4,792 on the basis of her experience only, but this is still $109 less than a male teacher with zero years of experience.

A more complex interaction between a dummy variable and the dependent variable can also be estimated. It may be that the dummy variable has more than a simple shift effect on the dependent variable, but also interacts with one or more of the other continuous independent variables. While not tested in the example above, it could be hypothesized that the impact of gender on salary was not a one-time shift, but impacted the value of additional years of experience on salary also. That is, female school teacher's salaries were discounted at the start, and further did not grow at the same rate from the effect of experience as for male school teachers. This would show up as a different slope for the relationship between total years of experience for males than for females. If this is so then females school teachers would not just start behind their male colleagues (as measured by the shift in the estimated regression line), but would fall further and further behind as time and experienced increased.

The graph below shows how this hypothesis can be tested with the use of dummy variables and an interaction variable.

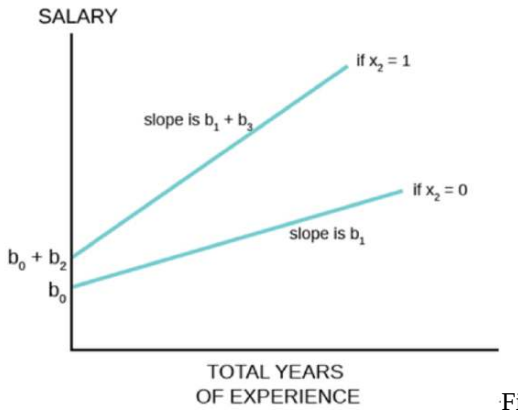


Figure 10.5.8

The estimating equation shows how the slope of X_1, the continuous random variable experience, contains two parts, b_1 and b_3:

$$\hat{Y} = b_0 + b_2X_2 + b_1X_1 + b_3X_2X_1$$

This occurs because of the new variable $X_2\ X_1$, called the interaction variable, was created to allow for an effect on the slope of X_1 from changes in X_2, the binary dummy variable. Note that when the dummy variable, $X_2 = 0$ the interaction variable has a value of 0, but when $X_2 = 1$ the interaction variable has a value of X_1. The coefficient b_3 is an estimate of the difference in the coefficient of X_1 when $X_2 = 1$ compared to when $X_2 = 0$. In the example of teacher's salaries, if there is a premium paid to male teachers that affects the rate of increase in salaries from experience, then the rate at which male teachers' salaries rises would be $b_1 + b_3$ and the rate at which female teachers' salaries rise would be simply b_1. This hypothesis can be tested with the hypothesis:

$$H_0 : \beta_3 = 0 | \beta_1 = 0, \beta_2 = 0$$

$$H_a : \beta_3 \neq 0 | \beta_1 \neq 0, \beta_2 \neq 0$$

This is a t-test using the test statistic for the parameter β_3. If we reject the null hypothesis that $\beta_3 = 0$ we conclude there is a difference between the rate of increase for the group for whom the value of the binary variable is set to 1, males in this example. This estimating equation can be combined with our earlier one that tested only a parallel shift in the estimated line. The earnings/experience functions in Figure 10.5.8 are drawn for this case with a shift in the earnings function and a difference in the slope of the function with respect to total years of experience.

Exercise 10.5.1

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 randomly selected student if you know the third exam score?

X (third exam score)	Y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Table 10.5.2

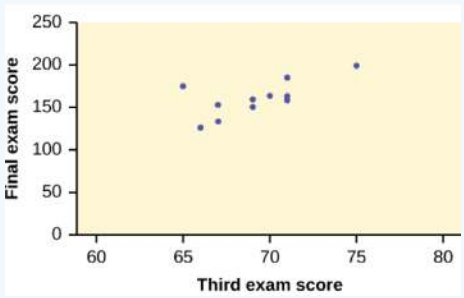


Figure 10.5.9 Scatter plot showing the scores on the final exam based on scores from the third exam.

Example 10.5.2

Recall Example 10.5.1 on the third exam and final exam scores.

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. Assume the coefficient for X was determined to be significantly different from zero.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores **(X-values)** range from 65 to 75. Since 73 is between the X variable values 65 and 75, we feel comfortable to substitute $X = 73$ into the equation. Then:

$$\hat{Y} = -173.51 + 4.83(73) = 179.08$$

We 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.

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Answer

a. 145.27

b. 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 have a confidence interval that may not be meaningful.)

To understand really how unreliable the prediction can be outside of the observed X values observed in the data, make the substitution $X = 90$ into the equation.

$\hat{Y} = -173.51 + 4.83(90) = 261.19$

The final-exam score is predicted to be 261.19. The largest the final-exam score can be is 200.

This page titled 10.5: The Regression Equation is shared under a CC BY license and was authored, remixed, and/or curated by .

10.6: How to Use Microsoft Excel® for Regression Analysis

This section of this chapter is here in recognition that what we are now asking requires much more than a quick calculation of a ratio or a square root. Indeed, the use of regression analysis was almost non- existent before the middle of the last century and did not really become a widely used tool until perhaps the late 1960's and early 1970's. Even then the computational ability of even the largest IBM machines is laughable by today's standards. In the early days programs were developed by the researchers and shared. There was no market for something called "software" and certainly nothing called "apps", an entrant into the market only a few years old.

With the advent of the personal computer and the explosion of a vital software market we have a number of regression and statistical analysis packages to choose from. Each has their merits. We have chosen Microsoft Excel because of the wide-spread availability both on college campuses and in the post-college market place. Stata is an alternative and has features that will be important for more advanced econometrics study if you choose to follow this path. Even more advanced packages exist, but typically require the analyst to do some significant amount of programing to conduct their analysis. The goal of this section is to demonstrate how to use Excel to run a regression and then to do so with an example of a simple version of a demand curve.

The first step to doing a regression using Excel is to load the program into your computer. If you have Excel you have the Analysis ToolPak although you may not have it activated. The program calls upon a significant amount of space so is not loaded automatically.

To activate the Analysis ToolPak follow these steps:

Click "File" > "Options" > "Add-ins" to bring up a menu of the add-in "ToolPaks". Select "Analysis ToolPak" and click "GO" next to "Manage: excel add-ins" near the bottom of the window. This will open a new window where you click "Analysis ToolPak" (make sure there is a green check mark in the box) and then click "OK". Now there should be an Analysis tab under the data menu. These steps are presented in the following screen shots.

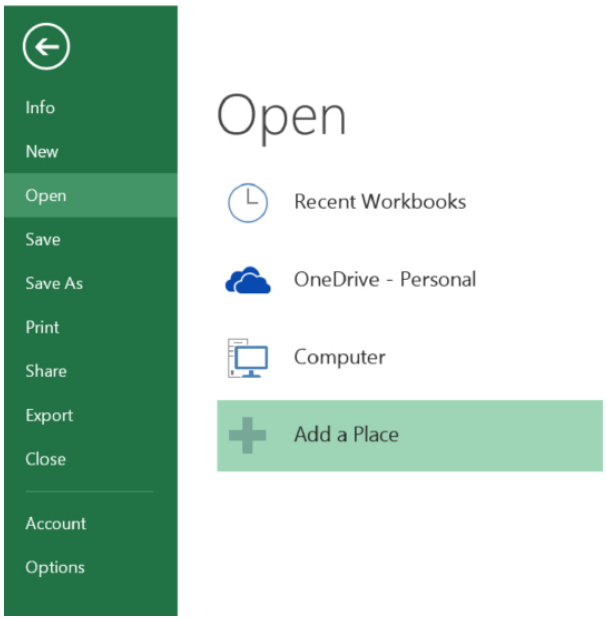


Figure 10.6.1

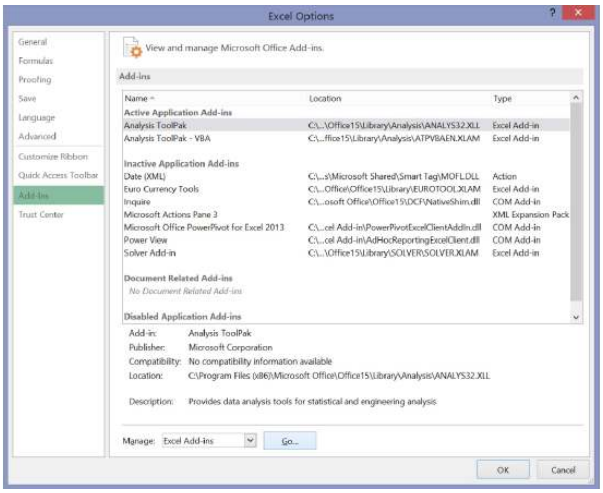


Figure 10.6.2

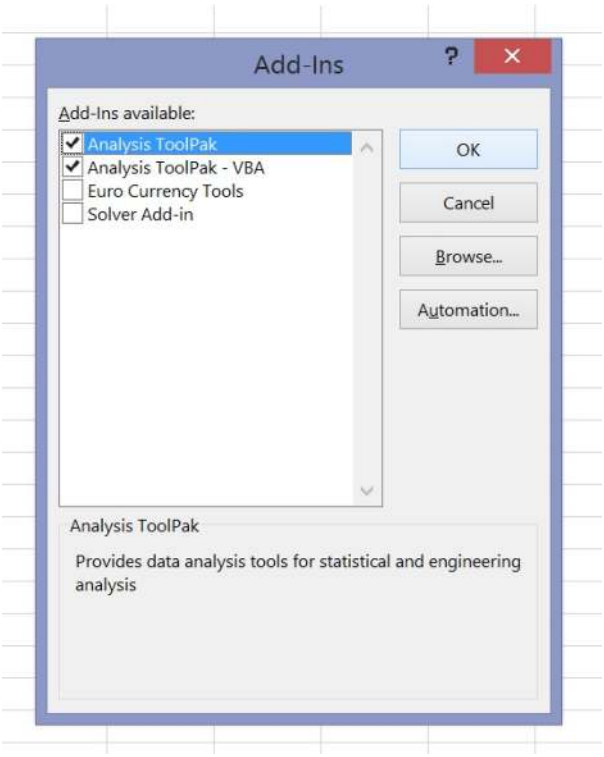


Figure 10.6.3

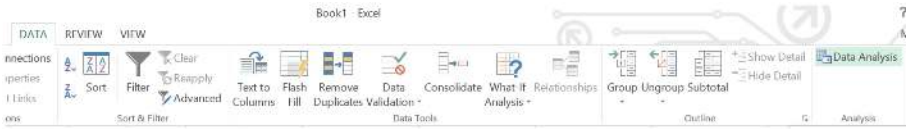


Figure 10.6.4

Click “Data” then “Data Analysis” and then click “Regression” and “OK”. Congratulations, you have made it to the regression window. The window asks for your inputs. Clicking the box next to the Y and X ranges will allow you to use the click and drag feature of Excel to select your input ranges. Excel has one odd quirk and that is the click and drop feature requires that the independent variables, the X variables, are all together, meaning that they form a single matrix. If your data are set up with the Y variable between two columns of X variables Excel will not allow you to use click and drag. As an example, say Column A and Column C are independent variables and Column B is the Y variable, the dependent variable. Excel will not allow you to click and

drop the data ranges. The solution is to move the column with the Y variable to column A and then you can click and drag. The same problem arises again if you want to run the regression with only some of the X variables. You will need to set up the matrix so all the X variables you wish to regress are in a tightly formed matrix. These steps are presented in the following scene shots.

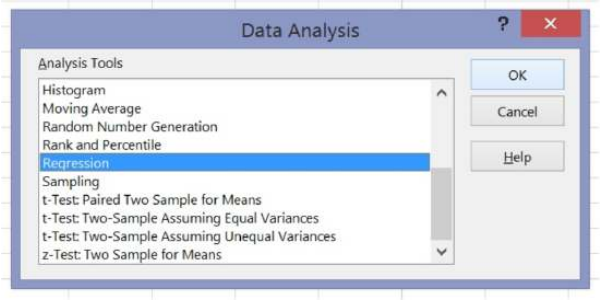


Figure 10.6.5

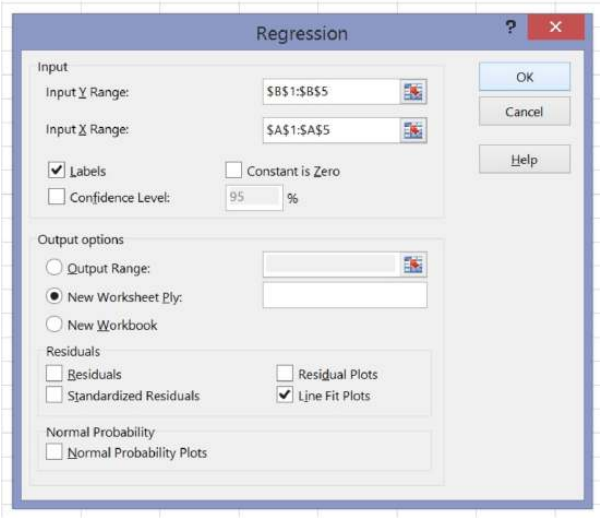


Figure 10.6.6

Once you have selected the data for your regression analysis and told Excel which one is the dependent variable (Y) and which ones are the independent valuables (X's), you have several choices as to the parameters and how the output will be displayed. Refer to screen shot Figure 10.6.6 under "Input" section. If you check the "labels" box the program will place the entry in the first column of each variable as its name in the output. You can enter an actual name, such as price or income in a demand analysis, in row one of the Excel spreadsheet for each variable and it will be displayed in the output.

The level of significance can also be set by the analyst. This will not change the calculated t-statistic, called t stat, but will alter the p-value for the calculated t-statistic. It will also alter the boundaries of the confidence intervals for the coefficients. A 95 percent confidence interval is always presented, but with a change in this you will also get other levels of confidence for the intervals.

Excel also will allow you to suppress the intercept. This forces the regression program to minimize the residual sum of squares under the condition that the estimated line must go through the origin. This is done in cases where there is no meaning in the model at some value other than zero, zero for the start of the line. An example is an economic production function that is a relationship between the number of units of an input, say hours of labor, and output. There is no meaning of positive output with zero workers.

Once the data are entered and the choices are made click OK and the results will be sent to a separate new worksheet by default. The output from Excel is presented in a way typical of other regression package programs. The first block of information gives the overall statistics of the regression: Multiple R, R Squared, and the R squared adjusted for degrees of freedom, which is the one you want to report. You also get the Standard error (of the estimate) and the number of observations in the regression.

The second block of information is titled ANOVA which stands for Analysis of Variance. Our interest in this section is the column marked F. This is the calculated F statistics for the null hypothesis that all of the coefficients are equal to zero verse the alternative

that at least one of the coefficients are not equal to zero. This hypothesis test was presented in Chapter 10.5 under "How Good is the Equation?" The next column gives the p-value for this test under the title "Significance F". If the p-value is less than say 0.05 (the calculated F-statistic is in the tail) we can say with 90% confidence that we reject the null hypotheses that all the coefficients are equal to zero. This is a good thing: it means that at least one of the coefficients is significantly different from zero thus do have an effect on the value of Y.

The last block of information contains the hypothesis tests for the individual coefficient. The estimated coefficients, the intercept and the slopes, are first listed and then each standard error (of the estimated coefficient) followed by the t stat (calculated Student's t-statistic for the null hypothesis that the coefficient is equal to zero). We compare the t stat and the critical value of the Student's t, dependent on the degrees of freedom, and determine if we have enough evidence to reject the null that the variable has no effect on Y. Remember that we have set up the null hypothesis as the status quo and our claim that we know what caused the Y to change is in the alternative hypothesis. We want to reject the status quo and substitute our version of the world, the alternative hypothesis. The next column contains the p-values for this hypothesis test followed by the estimated upper and lower bound of the confidence interval of the estimated slope parameter for various levels of confidence set by us at the beginning.

Estimating the Demand for Roses

Here is an example of using the Excel program to run a regression for a particular specific case: estimating the demand for roses. We are trying to estimate a demand curve, which from economic theory we expect certain variables affect how much of a good we buy. The relationship between the price of a good and the quantity demanded is the demand curve. Beyond that we have the demand function that includes other relevant variables: a person's income, the price of substitute goods, and perhaps other variables such as season of the year or the price of complimentary goods. Quantity demanded will be our Y variable, and Price of roses, Price of carnations and Income will be our independent variables, the X variables.

For all of these variables theory tells us the expected relationship. For the price of the good in question, roses, theory predicts an inverse relationship, the negatively sloped demand curve. Theory also predicts the relationship between the quantity demanded of one good, here roses, and the price of a substitute, carnations in this example. Theory predicts that this should be a positive or direct relationship; as the price of the substitute falls we substitute away from roses to the cheaper substitute, carnations. A reduction in the price of the substitute generates a reduction in demand for the good being analyzed, roses here. Reduction generates reduction is a positive relationship. For normal goods, theory also predicts a positive relationship; as our incomes rise we buy more of the good, roses. We expect these results because that is what is predicted by a hundred years of economic theory and research. Essentially we are testing these century-old hypotheses. The data gathered was determined by the model that is being tested. This should always be the case. One is not doing inferential statistics by throwing a mountain of data into a computer and asking the machine for a theory. Theory first, test follows.

These data here are national average prices and income is the nation's per capita personal income. Quantity demanded is total national annual sales of roses. These are annual time series data; we are tracking the rose market for the United States from 1984-2017, 33 observations.

Because of the quirky way Excel requires how the data are entered into the regression package it is best to have the independent variables, price of roses, price of carnations and income next to each other on the spreadsheet. Once your data are entered into the spreadsheet it is always good to look at the data. Examine the range, the means and the standard deviations. Use your understanding of descriptive statistics from the very first part of this course. In large data sets you will not be able to "scan" the data. The Analysis ToolPak makes it easy to get the range, mean, standard deviations and other parameters of the distributions. You can also quickly get the correlations among the variables. Examine for outliers. Review the history. Did something happen? Was here a labor strike, change in import fees, something that makes these observations unusual? Do not take the data without question. There may have been a typo somewhere, who knows without review.

Go to the regression window, enter the data and select 95% confidence level and click "OK". You can include the labels in the input range if you have put a title at the top of each column, but be sure to click the "labels" box on the main regression page if you do.

The regression output should show up automatically on a new worksheet.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.8560327					
R Square	0.732792					
Adjusted R Square	0.699391					
Standard Error	3629.3427					
Observations	33					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	577972629.2	2.89E+08	21.9392274	2.59893E-05	
Residual	29	210754050.4	13172128			
Total	32	788726679.5				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	183475.43	16791.81835	10.92648	7.89854E-09	147878.367	219072.5
Price of Roses	-1.7607	0.2982	-5.9043	5.20E-05	-2.4049	-1.1164
Price of Carnations	1.3397	0.5273	2.5407	0.0246	0.208	2.4789
Income (per capita)	3.0338	1.2308	2.464901	0.00886322	0.621432	5.4446

Figure 10.6.7

The first results presented is the R-Square, a measure of the strength of the correlation between Y and X_1, X_2, and X_3 taken as a group. Our R-square here of 0.699, adjusted for degrees of freedom, means that 70% of the variation in Y, demand for roses, can be explained by variations in X_1, X_2, and X_3, Price of roses, Price of carnations and Income. There is no statistical test to determine the "significance" of an R^2. Of course a higher R^2 is preferred, but it is really the significance of the coefficients that will determine the value of the theory being tested and which will become part of any policy discussion if they are demonstrated to be significantly different form zero.

Looking at the third panel of output we can write the equation as:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + e$$

where b_0 is the intercept, b_1 is the estimated coefficient on price of roses, and b_2 is the estimated coefficient on price of carnations, b_3 is the estimated effect of income and e is the error term. The equation is written in Roman letters indicating that these are the estimated values and not the population parameters, β's.

Our estimated equation is:

$$\begin{aligned} &\text{Quantity of roses sold } = 183, \\ &475 - 1.76 \text{ Price of roses } + 1.33 \text{ Price of carnations } + 3.03 \text{ Income} \end{aligned}$$

We first observe that the signs of the coefficients are as expected from theory. The demand curve is downward sloping with the negative sign for the price of roses. Further the signs of both the price of carnations and income coefficients are positive as would be expected from economic theory.

Interpreting the coefficients can tell us the magnitude of the impact of a change in each variable on the demand for roses. It is the ability to do this which makes regression analysis such a valuable tool. The estimated coefficients tell us that an increase the price of roses by one dollar will lead to a 1.76 reduction in the number roses purchased. The price of carnations seems to play an important role in the demand for roses as we see that increasing the price of carnations by one dollar would increase the demand for roses by 1.33 units as consumers would substitute away from the now more expensive carnations. Similarly, increasing per capita income by one dollar will lead to a 3.03 unit increase in roses purchased.

These results are in line with the predictions of economics theory with respect to all three variables included in this estimate of the demand for roses. It is important to have a theory first that predicts the significance or at least the direction of the coefficients. Without a theory to test, this research tool is not much more helpful than the correlation coefficients we learned about earlier.

We cannot stop there, however. We need to first check whether our coefficients are statistically significant from zero. We set up a hypothesis of:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

for all three coefficients in the regression. Recall from earlier that we will not be able to definitively say that our estimated b_1 is the actual real population of β_1, but rather only that with $(1-\alpha)\%$ level of confidence that we cannot reject the null hypothesis that our estimated β_1 is significantly different from zero. The analyst is making a claim that the price of roses causes an impact on quantity demanded. Indeed, that each of the included variables has an impact on the quantity of roses demanded. The claim is therefore in the alternative hypotheses. It will take a very large probability, 0.95 in this case, to overthrow the null hypothesis, the status quo, that $\beta = 0$. In all regression hypothesis tests the claim is in the alternative and the claim is that the theory has found a variable that has a significant impact on the Y variable.

The test statistic for this hypothesis follows the familiar standardizing formula which counts the number of standard deviations, t, that the estimated value of the parameter, b_1, is away from the hypothesized value, β_0, which is zero in this case:

$$t_{obs} = \frac{b_1 - \beta_0}{S_{b_1}}$$

The computer calculates this test statistic and presents it as "t stat". You can find this value to the right of the standard error of the coefficient estimate. The standard error of the coefficient for b_1 is s_{b_1} in the formula. To reach a conclusion we compare this observed test statistic with the critical value of the Student's t at degrees of freedom $n-3-1=29$, and alpha = 0.025 (5% significance level for a two-tailed test). Our t stat for b_1 is approximately 5.90 which is greater than 1.96 (the critical value we looked up in the t-table), so we reject our null hypothesis of no effect. We conclude that Price has a significant effect because the observed t-value is in the tail. We conduct the same test for b2 and b3. For each variable, we find that we reject the null hypothesis of no relationship because the observed t-statistics are in the tail for each case, that is, greater than the critical value. All variables in this regression have been determined to have a significant effect on the demand for roses.

These tests tell us whether or not an individual coefficient is significantly different from zero, but does not address the overall quality of the model. We have seen that the R squared adjusted for degrees of freedom indicates this model with these three variables explains 70% of the variation in quantity of roses demanded. We can also conduct a second test of the model taken as a whole. This is the F test presented in section 13.4 of this chapter. Because this is a multiple regression (more than one X), we use the F-test to determine if our coefficients collectively affect Y. The hypothesis is:

$$H_0 : \beta_1 = \beta_2 = ... = \beta i = 0$$

$$H_a : "\text{at least one of the} \beta_i \text{ is not equal to } 0"$$

Under the ANOVA section of the output we find the calculated F statistic for this hypotheses. For this example the F statistic is 21.9. Again, comparing the calculated F-statistic with the critical value given our desired level of significance and the degrees of freedom will allow us to reach a conclusion.

The best way to reach a conclusion for this statistical test is to use the p-value comparison rule. The p-value is the area in the tail, given the calculated F-statistic. In essence the computer is finding the F value in the table for us and calculating the p-value. In the Summary Output under "significance F" is this probability. For this example, it is calculated to be 2.6 X 10-5, or 2.6 then moving the decimal five places to the left. (.000026) This is an almost infinitesimal level of probability and is certainly less than our alpha level of .05 for a 5 percent level of significance.

By rejecting the null hypotheses we conclude that this specification of this model has validity because at least one of the estimated coefficients is significantly different from zero. Since F-observed is greater than F-critical, we reject H_0, meaning that X_1, X_2 and X_3 together have a significant effect on Y.

The development of computing machinery and the software useful for academic and business research has made it possible to answer questions that just a few years ago we could not even formulate. Data is available in electronic format and can be moved into place for analysis in ways and at speeds that were unimaginable a decade ago. The sheer magnitude of data sets that can today be used for research and analysis gives us a higher quality of results than in days past. Even with only an Excel spreadsheet we can conduct very high level research. This section gives you the tools to conduct some of this very interesting research with the only limit being your imagination.

10.6: How to Use Microsoft Excel® for Regression Analysis is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.7: Chapter 10 Key Terms

a is the symbol for the Y-Intercept

Sometimes written as b_0, because when writing the theoretical linear model β_0 is used to represent a coefficient for a population.

b is the symbol for Slope

The word coefficient will be used regularly for the slope, because it is a number that will always be next to the letter "X." It will be written as b_1 when a sample is used, and β_1 will be used with a population or when writing the theoretical linear model.

Bivariate

two variables are present in the model where one is the "cause" or independent variable and the other is the "effect" of dependent variable.

Linear

a model that takes data and regresses it into a straight line equation.

Multivariate

a system or model where more than one independent variable is being used to predict an outcome. There can only ever be one dependent variable, but there is no limit to the number of independent variables.

R^2 – Coefficient of Determination

This is a number between 0 and 1 that represents the percentage variation of the dependent variable that can be explained by the variation in the independent variable.

Residual or "error"

the value calculated from subtracting $Y_0 - \hat{Y}_0 = e_0$. The absolute value of a residual measures the vertical distance between the actual value of Y and the estimated value of Y that appears on the best-fit line.

r – Correlation Coefficient

A number between -1 and 1 that represents the strength and direction of the relationship between X and Y. The value for r will equal 1 or -1 only if all the plotted points form a perfectly straight line.

Sum of Squared Errors (SSE)

the calculated value from adding up all the squared residual terms. The hope is that this value is very small when creating a model.

X – the independent variable

This will sometimes be referred to as the "predictor" variable, because these values were measured in order to determine what possible outcomes could be predicted.

Y – the dependent variable

Also, using the letter "Y" represents actual values while $\hat{Y}$ represents predicted or estimated values. Predicted values will come from plugging in observed X values into a linear model.

10.7: Chapter 10 Key Terms is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.8: Chapter 10 Review

10.4 Linear Equations

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $\mathbf{y = mx + b}$, where m and b are constants, X is the independent variable, and Y is the dependent variable. In a statistical context, a linear equation is written in the form $Y = a + bX$, where a and b are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $Y = a + bX$, the constant b that multiplies the X variable (b is called a coefficient) is called the **slope**. The slope describes the rate of change between the independent and dependent variables. In other words, the slope describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $Y = a + bX$, the constant a is called the Y-intercept.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (Y) changes for every one unit increase in the independent (X) variable, on average. The Y-**intercept** is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

10.5 The Regression Equation

It is hoped that this discussion of regression analysis has demonstrated the tremendous potential value it has as a tool for testing models and helping to better understand the world around us. The regression model has its limitations, especially the requirement that the underlying relationship be approximately linear. To the extent that the true relationship is nonlinear it may be approximated with a linear relationship or nonlinear forms of transformations that can be estimated with linear techniques. Double logarithmic transformation of the data will provide an easy way to test this particular shape of the relationship. A reasonably good quadratic form (the shape of the total cost curve from Microeconomics Principles) can be generated by the equation:

$$Y = a + b_1X + b_2X^2$$

where the values of X are simply squared and put into the equation as a separate variable.

There is much more in the way of econometric "tricks" that can bypass some of the more troublesome assumptions of the general regression model. This statistical technique is so valuable that further study would provide any student significant, statistically significant, dividends.

10.8: Chapter 10 Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.9: Chapter 10 Homework

10.2 The Correlation Coefficient *r*

1. In order to have a correlation coefficient between traits A and B, it is necessary to have:

a. one group of subjects, some of whom possess characteristics of trait A, the remainder possessing those of trait B
b. measures of trait A on one group of subjects and of trait B on another group
c. two groups of subjects, one which could be classified as A or not A, the other as B or not B

2. Define the correlation coefficient and give a unique example of its use.

3. If the correlation between age of an auto and money spent for repairs is +.90.

a. 81% of the variation in the money spent for repairs is explained by the age of the auto
b. 81% of money spent for repairs is unexplained by the age of the auto
c. 90% of the money spent for repairs is explained by the age of the auto
d. none of the above

4. Suppose that college grade-point average and verbal portion of an IQ test had a correlation of .40. What percentage of the variance do these two have in common?

a. 20
b. 16
c. 40
d. 80

5. True or false? If false, explain why: The coefficient of determination can have values between -1 and +1.

6. True or False: Whenever r is calculated on the basis of a sample, the value which we obtain for r is only an estimate of the true correlation coefficient which we would obtain if we calculated it for the entire population.

7. Under a "scatter diagram" there is a notation that the coefficient of correlation is .10. What does this mean?

a. plus and minus 10% from the means includes about 68% of the cases
b. one-tenth of the variance of one variable is shared with the other variable
c. one-tenth of one variable is caused by the other variable
d. on a scale from -1 to +1, the degree of linear relationship between the two variables is +.10

8. The correlation coefficient for X and Y is known to be zero. We then can conclude that:

a. X and Y have standard distributions
b. the variances of X and Y are equal
c. there exists no relationship between X and Y
d. there exists no linear relationship between X and Y
e. none of these

9. What would you guess the value of the correlation coefficient to be for the pair of variables: "number of hours worked" and "number of units of work completed"?

a. Approximately 0.9
b. Approximately 0.4
c. Approximately 0.0
d. Approximately -0.4
e. Approximately -0.9

10. In a given group, the correlation between height measured in feet and weight measured in pounds is +.68. Which of the following would alter the value of r?

a. height is expressed centimeters.

b. weight is expressed in kilograms.
c. both of the above will affect r.
d. neither of the above changes will affect r.

10.3 Testing the Significance of the Correlation Coefficient

11. Use the dataset below to determine whether there is a significant correlation between the following monthly returns. Use the 95% confidence level.

Month	Apple Inc.	S&P 500 ETF	Southern California Edison
Jan	1	1	6
Feb	4	3	5
Mar	10	1	2
Apr	6	4	4
May	-13	-6	-3
Jun	13	6	1
Jul	8	2	5
Aug	-2	-2	3
Sep	7	1	0
Oct	11	2	-4
Nov	7	4	0
Dec	10	2	5

Table 10.9.1

a. Write the pair of hypotheses that would test whether there is a significant correlation between the monthly returns of Apple Inc. and S&P 500.
b. Calculate the relevant correlation coefficient r.
c. Test the above hypotheses using test statistics. Interpret your results.
d. Write the pair of hypotheses that would test whether there is a significant correlation between the monthly returns of Southern California Edison and S&P 500.
e. Calculate the relevant correlation coefficient r.
f. Test the above hypotheses using test statistics. Interpret your results.
g. Explain why there would be a difference between the results in part c. and part f.

12. The correlation between scores on a neuroticism test and scores on an anxiety test is high and positive. Therefore, we can conclude that:

a. anxiety causes neuroticism.
b. those who score low on one test tend to score high on the other.
c. those who score low on one test tend to score low on the other.
d. no prediction from one test to the other can be meaningfully made.

10.4 Linear Equations

13. True or False? If False, correct it: Suppose a 95% confidence interval for the slope β of the straight line regression of Y on X is given by $-3.5 < \beta < -0.5$. Then a two-tailed test of the hypothesis $H_0 : \beta = -1$ would result in rejection of H_0 at the 1% level of significance.

14. True or False: It is safer to interpret correlation coefficients as measures of association rather than causation because of the possibility of spurious correlation.

15. We are interested in finding the linear relation between the number of widgets purchased at one time and the cost per widget. The following data has been obtained:

X = Number of widgets purchased: 1, 3, 6, 10, 15

Y = Cost per widget (in dollars): 55, 52, 46, 32, 25

Suppose the regression line is $\hat{Y} = -2.5X + 60$. We compute the average price per widget if 30 are purchased and observe which of the following?

a. $\hat{Y} = 15 \text{ dollars}$; obviously, we are mistaken; the prediction $\hat{Y}$ is actually +15 dollars.
b. $\hat{Y} = 15 \text{ dollars}$, which seems reasonable judging by the data.
c. $\hat{Y} = -15 \text{ dollars}$, which is obvious nonsense. The regression line must be incorrect.
d. $\hat{Y} = -15 \text{ dollars}$, which is obvious nonsense. This reminds us that predicting Y outside the range of X values in our data is a very poor practice.

16. Discuss briefly the distinction between correlation and causality.

17. True or False: If r is close to +1 or -1, we shall say there is a strong correlation, with the tacit understanding that we are referring to a linear relationship and nothing else.

10.5 The Regression Equation

18. Suppose that you have at your disposal the information below for each of 30 drivers. Propose a model (including a very brief indication of symbols used to represent independent variables) to explain how miles per gallon vary from driver to driver on the basis of the factors measured.

Information:

1. miles driven per day
2. weight of car
3. number of cylinders in car
4. average speed
5. miles per gallon
6. number of passengers

19. Consider a sample least squares regression analysis between a dependent variable (Y) and an independent variable (X). A sample correlation coefficient of −1 (minus one) tells us that

a. there is no relationship between Y and X in the sample
b. there is no relationship between Y and X in the population
c. there is a perfect negative relationship between Y and X in the population
d. there is a perfect negative relationship between Y and X in the sample.

20. In correlational analysis, when the points scatter widely about the regression line, this means that the correlation is

a. negative.
b. low.
c. heterogeneous.
d. between two measures that are unreliable.

21. In a linear regression, why do we need to be concerned with the range of the independent (X) variable?

22. ABC International wants to explore the relationship between the yearly marketing expenses and sales revenues (in millions USD) (see table below).

Marketing Expenses	Sales Revenues
4	8
2	4
8	18
6	22
10	30
6	8

Table 10.9.2

a. Determine the correlation coefficient predicting yearly sales revenues from marketing expenses.
b. Write a pair of hypotheses that would test whether marketing expenses and sales revenues are significantly correlated.
c. Test the above hypotheses using test statistics at a 95% confidence level.
d. Test the above hypotheses using *p*-values at a 95% confidence level. Interpret the *p*-value.

23. An economist is interested in the possible influence of "Miracle Wheat" on the average yield of wheat in a district. To do so he fits a linear regression of average yield per year against year after introduction of "Miracle Wheat" for a ten year period.

The fitted trend line is $\hat{Y}_j = 80 + 1.5X_j$.

(Y_j: Average yield in j year after introduction)

(X_j: j year after introduction).

1. What is the estimated average yield for the fourth year after introduction?
2. Do you want to use this trend line to estimate yield for, say, 20 years after introduction? Why? What would your estimate be?

24. An interpretation of $r = 0.5$ is that the following part of the Y-variation is associated with which variation in X:

a. most
b. half
c. very little
d. one quarter
e. none of these

25. Which of the following values of r indicates the most accurate prediction of one variable from another?

a. $r = 1.18$
b. $r = -.77$
c. $r = .68$

26. Historically, returns on stocks and bonds are negatively correlated. Use the data in the table below to test whether this is true for a given investor across the past five months.

Month	Stock Return (%)	Bond Return (%)
1	4.2	1.0
2	-2.1	1.8
3	3.5	0.5
4	-1.2	2.0
5	0.0	1.2

Table 10.9.3

a. Determine the coefficient of the correlation between this investor's monthly returns on stocks vs. bonds.
b. Write the pair of hypotheses that would test whether monthly returns on stocks and bonds are negatively correlated.
c. Test the above hypotheses using test statistics at a 95% confidence level.
d. Test the above hypotheses using *p*-values at a 95% confidence level. Interpret the *p*-value.

27. A manager wants to examine whether the number of hours team leaders spend on employee coaching is related to their team's weekly performance score. The data from seven teams is reported in the table below. Use 99% confidence.

Team	Coaching Hours	Performance Score
A	2	65
B	5	75
C	3	70
D	6	65
E	4	80
F	2	75
G	0	65

Table 10.9.4

a. Determine the correlation coefficient here.
b. Write the pair of hypotheses that would test whether the number of hours team leaders spend on employee coaching is related to their team's weekly performance score.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.

28. A manager wants to examine whether working remotely (measured in number of remote hours per week, with other work time completed in-office) affects the number of tasks that employees complete. The data from five employees is reported in the table below. Use 5% error probability.

Employee	Remote Hours	# of Tasks Completed
A	16	45
B	8	38
C	8	30
D	8	42
E	4	28

Table 10.9.5

a. Determine the correlation coefficient here.
b. Write the pair of hypotheses being tested here.
c. Test the above hypotheses using test statistics.
d. Test the above hypotheses using *p*-values. Interpret the *p*-value.

10.9: Chapter 10 Homework is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.10: Chapter 10 Solutions

1. c

2. A measure of the degree to which variation of one variable is related to variation in one or more other variables. The most commonly used correlation coefficient indicates the degree to which variation in one variable is described by a straight line relation with another variable.

Suppose that sample information is available on family income and Years of schooling of the head of the household. A correlation coefficient = 0 would indicate no linear association at all between these two variables. A correlation of 1 would indicate perfect linear association (where all variation in family income could be associated with schooling and vice versa).

3. a. 81% of the variation in the money spent for repairs is explained by the age of the auto

4. b. 16

5. The coefficient of determination is $r2$ with $0 \leq r2 \leq 1$, since $-1 \leq r \leq 1$.

6. True

7. d. on a scale from -1 to +1, the degree of linear relationship between the two variables is +.10

8. d. there exists no linear relationship between X and Y

9. Approximately 0.9

10. d. neither of the above changes will affect r.

12. c. those who score low on one test tend to score low on the other.

13. False. Since $H_0 : \beta = -1$ would not be rejected at $\alpha = 0.05$, it would not be rejected at $\alpha = 0.01$.

14. True

15. d

16. Some variables seem to be related, so that knowing one variable's status allows us to predict the status of the other. This relationship can be measured and is called correlation. However, a high correlation between two variables in no way proves that a cause-and-effect relation exists between them. It is entirely possible that a third factor causes both variables to vary together.

17. True

19. d. there is a perfect negative relationship between Y and X in the sample.

20. b. low

21. The precision of the estimate of the Y variable depends on the range of the independent (X) variable explored. If we explore a very small range of the X variable, we won't be able to make much use of the regression. Also, extrapolation is not recommended.

23.

1. $80 + 1.5 * 4 = 86$
2. No. Most business statisticians would not want to extrapolate that far. If someone did, the estimate would be 110, but some other factors probably come into play with 20 years.

24. d. one quarter

25. b. $r = -.77$

10.10: Chapter 10 Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

11: Appendices

11.1: A - Statistical Tables

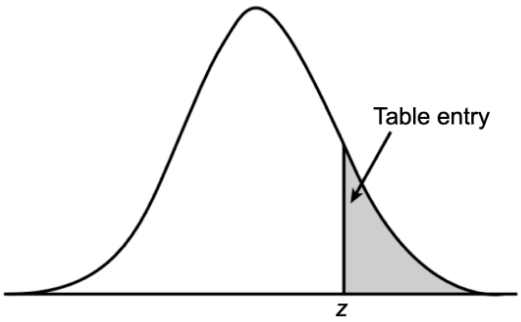
11.2: B - Mathematical Phrases, Symbols, and Formulas

11.3: C - Reporting Statistics in APA Style

This page titled 11: Appendices is shared under a CC BY license and was authored, remixed, and/or curated by .

Standard Normal Probability Distribution: *z* Table

Numerical entries represent the one-tailed probability that a standard normal random variable exceeds |*z*| (image modified from z-table.com).

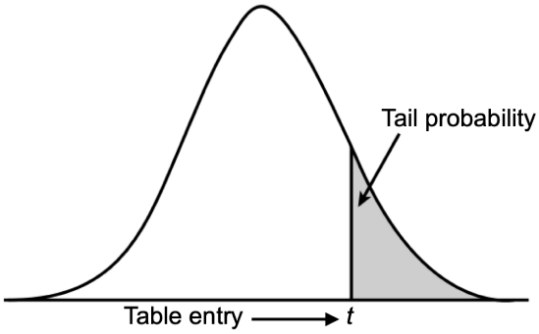


z	Second decimal place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048

z	Second decimal place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000233									
4.0	.0000317									
4.5	.00000340									
5.0	.000000287									

Student's *t* Distribution

The table shows the value of |*t*| that corresponds to the given one-tailed probabilities at varying degrees of freedom (*df*) (image modified from z-table.com).



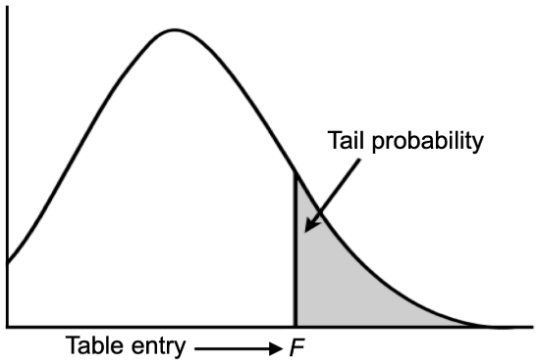
https://biz.libretexts.org/@go/page/79102

	Confidence Level (Two-Tailed)					
	80%	90%	95%	98%	99%	99.8%
	One-Tailed Probability					
df	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467

	Confidence Level (Two-Tailed)					
	80%	90%	95%	98%	99%	99.8%
	One-Tailed Probability					
df	0.10	0.05	0.025	0.01	0.005	0.001
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

F Distribution

The table shows the value of F that corresponds to the given right-tailed probabilities at varying degrees of freedom (df_1, df_2) (image modified from z-table.com).



	α = .05									
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91	249.05	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.20	2.01	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71

	α = .05									
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.02	1.83	1.61	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

https://biz.libretexts.org/@go/page/79102

df_2	α = .01									
	df_1									
	1	2	3	4	5	6	8	12	24	∞
1	4052.2	4999.5	5403.4	5624.6	5763.7	5859.0	5981.1	6106.3	6234.6	6365.9
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.75	10.93	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.17
14	8.86	6.52	5.56	5.04	4.70	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.79	3.46	3.08	2.65
18	8.29	6.01	5.09	4.58	4.25	4.02	3.71	3.37	3.00	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.93	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.77	4.26	3.94	3.71	3.41	3.07	2.70	2.26

df_2	α = .01									
	df_1									
	1	2	3	4	5	6	8	12	24	∞
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.59	2.13
27	7.68	5.49	4.60	4.11	3.79	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.05	3.73	3.50	3.20	2.87	2.50	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.67	2.29	1.81
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.51	2.19	1.79	1.00

https://biz.libretexts.org/@go/page/79102

df_2	α = .001									
	df_1									
	1	2	3	4	5	6	8	12	24	∞
1	405284	500000	540379	562500	576405	585937	598144	610668	623497	636619
2	998.50	999.00	999.17	999.25	999.30	999.33	999.37	999.42	999.46	999.50
3	167.03	148.50	141.11	137.10	134.58	132.85	130.62	128.32	125.93	123.50
4	74.14	61.25	56.18	53.44	51.71	50.53	49.00	47.41	45.77	44.05
5	47.18	37.12	33.20	31.09	29.75	28.83	27.65	26.42	25.13	23.78
6	35.51	27.00	23.70	21.92	20.80	20.03	19.03	17.99	16.90	15.75
7	29.25	21.69	18.77	17.20	16.21	15.52	14.63	13.71	12.73	11.69
8	25.41	18.49	15.83	14.39	13.48	12.86	12.05	11.19	10.30	9.34
9	22.86	16.39	13.90	12.56	11.71	11.13	10.37	9.57	8.72	7.81
10	21.04	14.91	12.55	11.28	10.48	9.93	9.20	8.45	7.64	6.76
11	19.69	13.81	11.56	10.35	9.58	9.05	8.35	7.63	6.85	6.00
12	18.64	12.97	10.80	9.63	8.89	8.38	7.71	7.00	6.25	5.42
13	17.82	12.31	10.21	9.07	8.35	7.86	7.21	6.52	5.78	4.97
14	17.14	11.78	9.73	8.62	7.92	7.44	6.80	6.13	5.41	4.60
15	16.59	11.34	9.34	8.25	7.57	7.09	6.47	5.81	5.10	4.31
16	16.12	10.97	9.01	7.94	7.27	6.80	6.19	5.55	4.85	4.06
17	15.72	10.66	8.73	7.68	7.02	6.56	5.96	5.32	4.63	3.85
18	15.38	10.39	8.49	7.46	6.81	6.35	5.76	5.13	4.45	3.67
19	15.08	10.16	8.28	7.27	6.62	6.18	5.59	4.97	4.29	3.52
20	14.82	9.95	8.10	7.10	6.46	6.02	5.44	4.82	4.15	3.38
21	14.59	9.77	7.94	6.95	6.32	5.88	5.31	4.70	4.03	3.26
22	14.38	9.61	7.80	6.81	6.19	5.76	5.19	4.58	3.92	3.15

df_2	α = .001									
	df_1									
	1	2	3	4	5	6	8	12	24	∞
23	14.20	9.47	7.67	6.70	6.08	5.65	5.09	4.48	3.82	3.05
24	14.03	9.34	7.55	6.59	5.98	5.55	4.99	4.39	3.74	2.97
25	13.88	9.22	7.45	6.49	5.89	5.46	4.91	4.31	3.66	2.89
26	13.74	9.12	7.36	6.41	5.80	5.38	4.83	4.24	3.59	2.82
27	13.61	9.02	7.27	6.33	5.73	5.31	4.76	4.17	3.52	2.75
28	13.50	8.93	7.19	6.25	5.66	5.24	4.69	4.11	3.46	2.70
29	13.39	8.85	7.12	6.19	5.59	5.18	4.64	4.05	3.41	2.64
30	13.29	8.77	7.05	6.12	5.53	5.12	4.58	4.00	3.36	2.59
40	12.61	8.25	6.59	5.70	5.13	4.73	4.21	3.64	3.01	2.23
60	11.97	7.77	6.17	5.31	4.76	4.37	3.86	3.32	2.69	1.90
120	11.38	7.32	5.78	4.95	4.42	4.04	3.55	3.02	2.40	1.56
∞	10.83	6.91	5.42	4.62	4.10	3.74	3.27	2.74	2.13	1.00

This page titled 11.1: A - Statistical Tables is shared under a CC BY license and was authored, remixed, and/or curated by .

11.2: B - Mathematical Phrases, Symbols, and Formulas

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4. The minimum of X is 4. X is no less than 4. X is greater than or equal to 4.	$X \geq 4$
X is at most 4. The maximum of X is 4. X is no more than 4. X is less than or equal to 4. X does not exceed 4.	$X \leq 4$
X is greater than 4. X is more than 4. X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$
X is 4. X is equal to 4. X is the same as to 4.	$X = 4$
X is not 4. X is not equal to 4. X is not the same as 4. X is different than 4.	$X \neq 4$

Symbols and Their Meanings

Chapter (First used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{}$	The square root of	same
Descriptive Statistics	Q_1	quartile one	the first quartile
Descriptive Statistics	Q_2	quartile two	the second quartile
Descriptive Statistics	Q_3	quartile three	the third quartile
Descriptive Statistics	IQR	interquartile range	Q_3–$Q_1 = IQR$
Descriptive Statistics	$\overline{x}$	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean
Descriptive Statistics	s	s	sample standard deviation
Descriptive Statistics	s^2	s squared	sample variance
Descriptive Statistics	σ	sigma	population standard deviation
Descriptive Statistics	σ^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{\}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring
Probability Topics	$P(A\|B)$	probability of A given B	probability of A occurring given B has occurred
Probability Topics	$P(A \cup B)$	probability of A or B	probability of A or B or both occurring
Probability Topics	$P(A \cap B)$	probability of A and B	probability of both A and B occurring (same time)
Probability Topics	A'	A-prime; complement of A	complement of A; not A
Probability Topics	$P(A')$	probability of the complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P(G_1)$	probability of green on first pick	same
The Normal Distribution	N	normal distribution	same
The Normal Distribution	z	z-score	same
The Normal Distribution	Z	standard normal distribution	same
The Central Limit Theorem	$\overline{x}$	x-bar	the random variable x-bar
The Central Limit Theorem	$\mu_{\overline{x}}$	mean of x-bars	the average of x-bars
The Central Limit Theorem	$\sigma_{\overline{x}}$	standard deviation of x-bars	same

Chapter (First used)	Symbol	Spoken	Meaning
Confidence Intervals	CL	confidence level	same
Confidence Intervals	CI	confidence interval	same
Confidence Intervals	EBM	error bound for a mean	same
Confidence Intervals	EBP	error bound for a proportion	same
Confidence Intervals	t	Student's t-distribution	same
Confidence Intervals	df	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	Student's t with $\alpha/2$ area in each tail	same
Confidence Intervals	P'	P-prime	sample proportion of success or interest
Hypothesis Testing	H_0	H-naught, H-sub-0	null hypothesis
Hypothesis Testing	H_a	H-a, H-sub a	alternative (or research) hypothesis
Hypothesis Testing	H_1	H-1, H-sub 1	alternative (or research) hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{x}_1 - \overline{x}_2$	x1-bar minus x2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	mu-1 minus mu-2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	P1-prime minus P2-prime	difference in sample proportions
Hypothesis Testing	$P_1 - P_2$	P1 minus P2	difference in population proportions
Linear Regression and Correlation	$Y = a + bX$	Y equals a plus b-X	equation of a straight line
Linear Regression and Correlation	$\hat{Y}$	Y-hat	estimated value of Y
Linear Regression and Correlation	r	sample correlation coefficient	same
Linear Regression and Correlation	ε	error term for a regression line	same
Linear Regression and Correlation	SSE	Sum of Squared Errors	same
F-Distribution and ANOVA	F	F-ratio	F-ratio

Formulas

Symbols you must know		
Population		**Sample**
N	Size	n
μ	Mean	$\overline{x}$
σ^2	Variance	s^2
σ	Standard deviation	s
P	Proportion	P'
Single data set formulae		
Population		**Sample**
$Q_3 = \frac{3(N+1)}{4}, Q_1 = \frac{(N+1)}{4}$	Inter-quartile range $IQR = Q_3 - Q_1$	$Q_3 = \frac{3(n+1)}{4}, Q_1 = \frac{(n+1)}{4}$
$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \cdot f_i$	Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot f_i$

Basic probability rules	
$P(A \cap B) = P(A\|B) \cdot P(B)$	Multiplication rule
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Addition rule
$P(A \cap B) = P(A) \cdot P(B)$ or $P(A\|B) = P(A)$	Independence test

The following formulae require the use of the z, t, or F tables.	
$z = \frac{x-\mu}{\sigma}$	z-transformation for normal distribution
Test statistics	**Confidence intervals** [bracketed symbols equal margin of error] (subscripts denote locations on respective distribution tables)
$z_{obs} = \frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	*Interval for the population mean when sigma is known* $\overline{x} \pm \left[z_{(\alpha/2)} \frac{\sigma}{\sqrt{n}}\right]$
$z_{obs} = \frac{\overline{x}-\mu_0}{\frac{s}{\sqrt{n}}}$	*Interval for the population mean when sigma is unknown and* $n > 100$ $\overline{x} \pm \left[z_{(\alpha/2)} \frac{s}{\sqrt{n}}\right]$
$t_{obs} = \frac{\overline{x}-\mu_0}{\frac{s}{\sqrt{n}}}$	*Interval for the population mean when sigma is unknown and* $n < 100$ $\overline{x} \pm \left[t_{(n-1),(\alpha/2)} \frac{s}{\sqrt{n}}\right]$
$z_{obs} = \frac{P'-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$	*Interval for the population proportion* $P' \pm \left[z_{(\alpha/2)} \sqrt{\frac{P'(1-P')}{n}}\right]$
$t_{obs} = \frac{\overline{x}_d-\mu_d}{\frac{s_d}{\sqrt{n}}}$	*Interval for difference between two means with matched pairs* $\overline{x}_d \pm \left[t_{(n-1),(\alpha/2)} \frac{s_d}{\sqrt{n}}\right]$ *where* s_d *is the deviation of the differences*
$z_{obs} = \frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	*Interval for difference between two independent means when* $n > 100$ $(\overline{x}_1 - \overline{x}_2) \pm \left[z_{(\alpha/2)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$
$z_{obs} = \frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	*Interval for difference between two independent means when* $n < 100$ $(\overline{x}_1 - \overline{x}_2) \pm \left[t_{(n_1+n_2-2),(\alpha/2)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$
	Interval for difference between two population proportions $(P_1' - P_2') \pm \left[z_{(\alpha/2)} \sqrt{\frac{P_1'(1-P_1')}{n_1} + \frac{P_2'(1-P_2')}{n_2}}\right]$

Simple linear regression formulae for $Y = a + b(X)$

$r_{XY} = \frac{\sum(X_i - \overline{X}) * (Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 * \sum (Y_i - \overline{Y})^2}}$ $r_{XY} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sqrt{\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right] * \left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right]}}$	Correlation coefficient
$b_1 = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2}$ $b_1 = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$ $b_1 = r_{XY}\left(\frac{s_Y}{s_X}\right)$	Coefficient b (or b_1, slope)

$b_0 = \overline{Y} - b_1\overline{X}$	Y-intercept (a, or b_0)
$s_e^2 = \frac{\Sigma\left(Y_i - \hat{Y}_i\right)^2}{n-k} = \frac{\sum_{i=1}^{n} e_i^2}{n-k}$	Estimate of the error variance
$s_b = \frac{s_e^2}{\sqrt{\left(X_i - \overline{X}\right)^2}} = \frac{s_e^2}{(n-1)s_X^2}$	Standard error for coefficient b
$t_{obs} = \frac{b-\beta_0}{s_b}$	Hypothesis test for coefficient β
$b \pm [t_{n-2,\alpha/2} s_b]$	Interval for coefficient β
$\hat{Y} \pm \left[t_{\alpha/2} * s_e \left(\sqrt{\frac{1}{n} + \frac{\left(X_p - \overline{X}\right)^2}{s_X}}\right)\right]$	Interval for expected value of Y
$\hat{Y} \pm \left[t_{\alpha/2} * s_e \left(\sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \overline{X}\right)^2}{s_X}}\right)\right]$	Prediction interval for an individual Y
ANOVA formulae	
$SS_R = n_1 (\bar{x}_1 - \bar{x})^2 + \cdots + n_g (\bar{x}_g - \bar{x})^2$	Sum of squares regression
$SS_E = (n_1 - 1) s_1^2 + \cdots + (n_g - 1) s_g^2$	Sum of squares error
$SS_T = SS_R + SS_E$	Sum of squares total
$R^2 = \frac{SS_R}{SS_T}$	Coefficient of determination

The following is the breakdown of a one-way ANOVA table for linear regression.

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F-ratio
Regression	$n_1 (\bar{x}_1 - \bar{x})^2 + \cdots + n_g (\bar{x}_g - \bar{x})^2$	1 or $g-1$	$MSR = \frac{SS_R}{df_R}$	$F = \frac{MS_R}{MS_E}$
Error	$(n_1 - 1) s_1^2 + \cdots + (n_g - 1) s_g^2$	$n-g$	$MSE = \frac{SS_E}{df_E}$	
Total	$SS_R + SS_E$	$n-1$		

This page titled 11.2: B - Mathematical Phrases, Symbols, and Formulas is shared under a CC BY license and was authored, remixed, and/or curated by .

11.3: C - Reporting Statistics in APA Style

Reporting Statistics in APA Style

APA style, much like any style guide, requires attention to detail (e.g., font, spaces, italics). The good news is that learning these guidelines now will help with your other Management coursework – APA style is required throughout this major.

Statistics will mainly be reported in the Results section when you write full-length research papers (for example, in Research Methods/MGT 301), but a few statistics may also be reported in the Method section of such papers (for example, describing the average age of your sample, along with the standard deviation).

General Guidelines

1. Choose a font and font size that is easily readable, and use this same font throughout your entire piece of work. The most common choice among APA style works is 12-point Times New Roman font. Other good options include 11-point Calibri, Arial, or Georgia. (And, always check specific class requirements: some instructors require a particular font and font size.)

2. All statistical symbols that are not Greek letters must be italicized (*M*, *SD*, *t*, *F*, *p*, …) each time you use them.

- However, parentheses and the numbers you report should NOT be italicized.
- And because CI is simply the acronym for the phrase "confidence interval," the letters CI will never be italicized.

3. Every equals sign (=) must have a space before and after it (the same holds for other mathematical operators, such as >, <).

- It helps to think of the equals sign as a representation of the word "equals" – just like you would have a space before and after each word, you must have a space before and after the symbol.
- For example, reporting ***M*=4.52** is not APA style, but ***M* = 4.52** is.

4. Means, *SD*s, test statistics, CIs and *p* values should be rounded to two decimal places.

- Report the exact *p* value to two or three decimal places, such as: ***p* = .03** or ***p* = .032.**
- If your exact *p* value is less than .001, it is conventional to report this as $p < .001$.

5. Put a zero before a decimal point when a number is less than 1 but that statistic can exceed 1 (*M*, *SD*, *t*, *F*...). Do not use a zero before a decimal when the statistic cannot be greater than 1 (proportion, *r*, *p*, *a* …).

- Report the test statistics with a zero before the decimal point (for example, ***t* = 0.25** is correct, but ***t* = .32** is not).
- Report the *p*-value or the correlation coefficient <u>without</u> a zero before the decimal point (for example, ***p* = .32** or ***r* = .32** is correct, but ***p* = 0.32** or ***r* = 0.32** is not).

6. Do not explain how or why you used a certain test unless it is unusual (assume your audience knows basic statistics).

7. Describe the finding in words, then support with appropriate statistics – for example:

- **There was a significant difference between the starting salaries of male and female employees, *t*(39) = 7.03, *p* = .02, 95% CI = [223.05, 530.61].**
- And add more details and supporting statistics where needed – for example, in the above statement, we are not yet clear on what group was higher or lower than the other (only have stated that there was a "difference"), so this needs to be added in another sentence:
  - **Specifically, male employees earned about $300 more per month on average (*M* = $903.31, *SD* = $156.67) than female employees (*M* = $595.75, *SD* = $133.22).**
  - You could also combine these two types of statements into one sentence – for example: **Male employees earned significantly more per month on average (*M* = $903.31, *SD* = $156.67) than female employees (*M* = $595.75, *SD* = $133.22), *t*(39) = 7.03, *p* = .02, 95% CI = [223.05, 530.61].**

Reporting Results of Specific Analyses

Report the appropriate statistics according to the type of statistical analyses you conduct (see general guideline #5 above).

One-sample *t*-test

Format your results by stating your finding and mentioning the values of the observed *t*-statistic, degrees of freedom and *p*-values using the general format *t*(df) = ____, *p* = ____

- Example: **The 125 subjects in our study had a mean age of 27.4 (*SD* = 1.8), and they were significantly older than the university norm of 19.5 years, *t*(124) = 2.65, *p* = .04, 95% CI = [0.2, 15.6].**

Independent-samples t-test

Format your results by stating your finding and mentioning the values of the observed *t*-statistic, degrees of freedom and *p*-values using the general format *t*(df) = ____, *p* = ____. If you find a significant difference between groups, indicate the direction of difference (which group has the higher mean?).

- Example: **Results indicate a significant preference for apple pie (*M* = 3.45, *SD* = 1.11) over cherry pie (*M* = 3.00, *SD* = .80), *t*(15) = 4.00, *p* = .001.**
- Example: **There was no significant difference between the severity of injuries of female athletes (*M* = 43.00, *SD* = 19.83) compared to male athletes (*M* = 41.95, *SD* = 16.81), *t*(88) = 0.25, *p* = .80.**
- Note: **See above (#7) for an example reporting a CI comparing two groups.**

Dependent-samples t-test

Format your results by stating your finding and mentioning the values of the observed *t*-statistic, degrees of freedom and *p*-values using the general format *t*(df) = ____, *p* = ____. If you find a significant difference between groups, indicate the direction of difference (which group has the higher mean?).

- Example: **The 228 college students had an average difference of -4.8 (*SD* = 5.5) between their college and high-school GPA scores, indicating a significant decrease in their college GPA scores, *t*(24) = -4.36, *p* = .003.**
- Example: **Children's second grade reading speed scores were typically higher (*M* = 58.09, *SD* = 34.67) than their first grade reading speed scores (*M* = 46.09, *SD* = 32.68), *t*(21) = -4.21, *p* < .001.**
- Note: **See above (#7) for an example reporting a CI comparing two groups.**

ANOVA

We report ANOVA results like *t*-tests, but with two degrees-of-freedom numbers and *F* statistic rounded off to 2 decimal places, in the format *F*(df between, df within) = _____, *p* =____.

- Remember: If the overall *F*-test is significant, there is a difference between at least one pair of group averages, but ANOVA does not tell which pair. You need to look at the results of the *post hoc t*-tests to find the two or more groups that are significantly different.
- If the overall F test is significant, describe the *post hoc t*-test results. Which pairs of group averages are significantly different? Report their *M*, *SD*, *t* and *p*-values to describe the differences (which group within each pair has the higher mean?).
- Example: **One-way analysis of variance showed that the average injury severity differed significantly across the various superhero costumes, *F*(3, 86) = 4.97, *p* = .003. Specifically, post-hoc comparisons show that children wearing Superman costumes had significantly more severe injuries (*M* = 54.17, *SD* = 17.82) than children in Ninja Turtle costumes (*M* = 34.42, *SD* = 17.85).**

Correlation

Correlations are reported using the general format *r*(df) = _____, *p* =____. Always mention the direction (positive, negative) and the strength of correlation (weak, moderate, strong).

- Example: **Hours spent studying and GPA were strongly positively correlated, *r*(123) = .61, *p* = .011.**
- Example: **Hours spent playing video games and GPA were moderately negatively correlated, *r*(123) = .32, *p* = .041.**

- Example: **There was a significant negative association between a child's age and severity of their injuries, $r(88) = -.73$, $p < .001$.**

Regression

Regression results are reported using the general format b = ____, t(df) = _____, p =____.

- Example: **As child's age increases by one year, the injury severity is predicted to decrease by 5.35 units, $t(88) = -9.88$, $p < .001$.** Or…. **There is a significant negative association between a child's age and injury severity, $b = -5.35$, $t(88) = -9.88$, $p < .001$.**
- You may also choose to report the percentage of variance explained along with the corresponding F test. For example: **Social support explained a significant proportion of variance in depression scores, $R^2 = .12$, $F(1, 225) = 42.64$, $p < .001$.** Or…. **Children's ages accounted for 52.6% of the variability of their reading speed score, $F(1, 88) = 97.65$, $p < .001$.**

11.3: C - Reporting Statistics in APA Style is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

Index

A

alternative hypothesis
- 7.2: Null and Alternative Hypotheses
- 7.4: Distribution Needed for Hypothesis Testing

average
- 1.2: Definitions of Statistics, Probability, and Key Terms

B

balanced design
- 9.3: The F-Distribution and the F-Ratio

bar graph
- 1.3: Data, Sampling, and Variation in Data and Sampling
- 2.2: Display Data

binomial distribution
- 6.4: A Confidence Interval for A Population Proportion
- 7.5: Full Hypothesis Test Examples

bivariate
- 10.2: The Correlation Coefficient r

blinding
- 1.5: Experimental Design and Ethics

C

categorical variables
- 1.2: Definitions of Statistics, Probability, and Key Terms

central limit theorem
- 5.1: Introduction to the Central Limit Theorem

coefficient of determination
- 10.5: The Regression Equation

coefficient of multiple determination
- 10.5: The Regression Equation

Cohen's d
- 8.3: Cohen's Standards for Small, Medium, and Large Effect Sizes

complement
- 3.2: Probability Terminology

conditional probability
- 3.2: Probability Terminology

Confidence Interval
- 6.1: Introduction
- 6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size
- 6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

confidence intervals
- 6.4: A Confidence Interval for A Population Proportion
- 7.1: Introduction to Hypothesis Testing

confidence level
- 6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size

contingency table
- 3.5: Contingency Tables and Probability Trees

continuous
- 1.3: Data, Sampling, and Variation in Data and Sampling

control group
- 1.5: Experimental Design and Ethics

correlation coefficient
- 10.2: The Correlation Coefficient r

critical values
- 7.4: Distribution Needed for Hypothesis Testing

cumulative relative frequency
- 1.4: Levels of Measurement

D

data
- 1.2: Definitions of Statistics, Probability, and Key Terms

degrees of freedom
- 6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

degrees of freedom (df)
- 8.2: Comparing Two Independent Population Means

dependent variable
- 1.5: Experimental Design and Ethics

descriptive statistics
- 1.2: Definitions of Statistics, Probability, and Key Terms

discrete
- 1.3: Data, Sampling, and Variation in Data and Sampling

E

empirical rule
- 4.2: The Standard Normal Distribution
- 6.1: Introduction

equal standard deviations
- 9.2: One-Way ANOVA

equally likely
- 3.2: Probability Terminology

error bound mean
- 6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size

estimate of the error variance
- 10.5: The Regression Equation

event
- 3.2: Probability Terminology

experiment
- 3.2: Probability Terminology

experimental unit
- 1.5: Experimental Design and Ethics

explanatory variable
- 1.5: Experimental Design and Ethics

F

F distribution
- 9.3: The F-Distribution and the F-Ratio

f ratio
- 9.3: The F-Distribution and the F-Ratio

fair
- 3.2: Probability Terminology

first moment
- 2.6: Skewness and the Mean, Median, and Mode

first quartile
- 2.3: Measures of the Location of the Data

frequency
- 1.4: Levels of Measurement
- 2.2: Display Data

H

histogram
- 2.2: Display Data

hypotheses
- 7.2: Null and Alternative Hypotheses

hypothesis test
- 7.5: Full Hypothesis Test Examples

hypothesis testing
- 7.1: Introduction to Hypothesis Testing

I

independent
- 3.3: Independent and Mutually Exclusive Events
- 3.4: Two Basic Rules of Probability

independent groups
- 8.1: Introduction

independent variable
- 1.5: Experimental Design and Ethics

inferential statistics
- 1.2: Definitions of Statistics, Probability, and Key Terms
- 6.1: Introduction

interquartile range
- 2.3: Measures of the Location of the Data

interval scale
- 1.4: Levels of Measurement

L

Law of Large Numbers
- 3.2: Probability Terminology
- 5.3: Using the Central Limit Theorem

level of measurement
- 1.4: Levels of Measurement

line graph
- 2.2: Display Data

lurking variables
- 1.5: Experimental Design and Ethics

M

matched pairs
- 8.1: Introduction

mean
- 1.2: Definitions of Statistics, Probability, and Key Terms
- 2.4: Measures of the Center of the Data

mean square
- 9.3: The F-Distribution and the F-Ratio

median
- 2.3: Measures of the Location of the Data
- 2.4: Measures of the Center of the Data

mode
- 2.4: Measures of the Center of the Data

multiple correlation coefficient
- 10.5: The Regression Equation

Multiplication Rule
- 3.4: Two Basic Rules of Probability

multivariate
10.2: The Correlation Coefficient r
mutually exclusive
3.3: Independent and Mutually Exclusive Events
3.4: Two Basic Rules of Probability

N

nominal scale
1.4: Levels of Measurement
normal distribution
6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case
7.4: Distribution Needed for Hypothesis Testing
null hypothesis
7.2: Null and Alternative Hypotheses
7.4: Distribution Needed for Hypothesis Testing
numerical variables
1.2: Definitions of Statistics, Probability, and Key Terms

O

ordinal scale
1.4: Levels of Measurement
outcome
3.2: Probability Terminology
outlier
2.2: Display Data
2.3: Measures of the Location of the Data

P

paired data set
2.2: Display Data
parameter
1.2: Definitions of Statistics, Probability, and Key Terms
6.1: Introduction
Pareto chart
1.3: Data, Sampling, and Variation in Data and Sampling
Pearson
1.2: Definitions of Statistics, Probability, and Key Terms
percentiles
2.3: Measures of the Location of the Data
pie chart
1.3: Data, Sampling, and Variation in Data and Sampling
placebo
1.5: Experimental Design and Ethics
point estimate
6.1: Introduction
population
1.2: Definitions of Statistics, Probability, and Key Terms
1.3: Data, Sampling, and Variation in Data and Sampling
power of the test
7.3: Outcomes and Type I and Type II Errors
preset or preconceived bfα
7.4: Distribution Needed for Hypothesis Testing
probability
1.2: Definitions of Statistics, Probability, and Key Terms
3.2: Probability Terminology
proportion
1.2: Definitions of Statistics, Probability, and Key Terms

Q

Qualitative Data
1.3: Data, Sampling, and Variation in Data and Sampling
quantitative continuous data
1.3: Data, Sampling, and Variation in Data and Sampling
Quantitative Data
1.3: Data, Sampling, and Variation in Data and Sampling
quantitative discrete data
1.3: Data, Sampling, and Variation in Data and Sampling
quartiles
2.3: Measures of the Location of the Data

R

random assignment
1.5: Experimental Design and Ethics
random variable
8.2: Comparing Two Independent Population Means
ratio scale
1.4: Levels of Measurement
Regression Equation
10.5: The Regression Equation
relative frequency
1.4: Levels of Measurement
2.2: Display Data
replacement
3.3: Independent and Mutually Exclusive Events
representative sample
1.2: Definitions of Statistics, Probability, and Key Terms
response variable
1.5: Experimental Design and Ethics

S

sample
1.2: Definitions of Statistics, Probability, and Key Terms
sample space
3.2: Probability Terminology
3.4: Two Basic Rules of Probability
3.5: Contingency Tables and Probability Trees
samples
1.3: Data, Sampling, and Variation in Data and Sampling
sampling
1.2: Definitions of Statistics, Probability, and Key Terms
second moment
2.6: Skewness and the Mean, Median, and Mode
significance level
7.4: Distribution Needed for Hypothesis Testing
skew
2.6: Skewness and the Mean, Median, and Mode
Standard Deviation
2.7: Measures of the Spread of the Data
6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case
7.4: Distribution Needed for Hypothesis Testing
standard error
8.2: Comparing Two Independent Population Means
standard error of the estimate
10.5: The Regression Equation
standard normal distribution
4.2: The Standard Normal Distribution
standardizing formula
4.3: Using the Normal Distribution
statistic
1.2: Definitions of Statistics, Probability, and Key Terms
statistics
1.2: Definitions of Statistics, Probability, and Key Terms
Stemplots
2.2: Display Data
sum of squared errors (SSE)
10.5: The Regression Equation
sum of squares
9.3: The F-Distribution and the F-Ratio

T

test statistic
7.4: Distribution Needed for Hypothesis Testing
the central limit theorem
5.2: The Central Limit Theorem for Sample Means
third quartile
2.3: Measures of the Location of the Data
treatments
1.5: Experimental Design and Ethics
type I
7.4: Distribution Needed for Hypothesis Testing
type I error
7.3: Outcomes and Type I and Type II Errors
type II error
7.3: Outcomes and Type I and Type II Errors

V

variable
1.2: Definitions of Statistics, Probability, and Key Terms
variance
2.7: Measures of the Spread of the Data
variance between samples
9.3: The F-Distribution and the F-Ratio
variance within samples
9.3: The F-Distribution and the F-Ratio
variances
9.2: One-Way ANOVA
variation
1.3: Data, Sampling, and Variation in Data and Sampling

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: MGT 235: Introductory Business Statistics (3rd edition)

Webpages: 127

All licenses found:

- CC BY 4.0: 55.1% (70 pages)
- Undeclared: 44.9% (57 pages)

By Page

- MGT 235: Introductory Business Statistics (3rd edition) - *CC BY 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Welcome to MGT 235! - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Sampling and Data - *CC BY 4.0*
    - 1.1: Introduction - *Undeclared*
    - 1.2: Definitions of Statistics, Probability, and Key Terms - *CC BY 4.0*
    - 1.3: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 1.4: Levels of Measurement - *CC BY 4.0*
    - 1.5: Experimental Design and Ethics - *CC BY 4.0*
    - 1.6: Chapter 1 Key Terms - *CC BY 4.0*
    - 1.7: Chapter 1 Review - *CC BY 4.0*
    - 1.8: Chapter 1 Homework - *Undeclared*
    - 1.9: Chapter 1 Solutions - *Undeclared*
    - 1.10: Chapter 1 References - *CC BY 4.0*
  - 2: Descriptive Statistics - *CC BY 4.0*
    - 2.1: Introduction - *Undeclared*
    - 2.2: Display Data - *CC BY 4.0*
    - 2.3: Measures of the Location of the Data - *CC BY 4.0*
    - 2.4: Measures of the Center of the Data - *CC BY 4.0*
    - 2.5: Sigma Notation and Calculating the Arithmetic Mean - *CC BY 4.0*
    - 2.6: Skewness and the Mean, Median, and Mode - *CC BY 4.0*
    - 2.7: Measures of the Spread of the Data - *CC BY 4.0*
    - 2.8: Chapter 2 Key Terms - *CC BY 4.0*
    - 2.9: Chapter 2 Review - *CC BY 4.0*
    - 2.10: Chapter 2 Formula Review - *CC BY 4.0*
    - 2.11: Chapter 2 Homework - *CC BY 4.0*
    - 2.12: Chapter 2 Solutions - *CC BY 4.0*
  - 3: Probability - *CC BY 4.0*
    - 3.1: Introduction to Probability - *CC BY 4.0*
    - 3.2: Probability Terminology - *CC BY 4.0*
    - 3.3: Independent and Mutually Exclusive Events - *CC BY 4.0*
    - 3.4: Two Basic Rules of Probability - *CC BY 4.0*
    - 3.5: Contingency Tables and Probability Trees - *CC BY 4.0*
    - 3.6: Chapter 3 Key Terms - *Undeclared*
    - 3.7: Chapter 3 Review - *Undeclared*
    - 3.8: Chapter 3 Formula Review - *Undeclared*
    - 3.9: Chapter 3 Homework - *Undeclared*
    - 3.10: Chapter 3 Solutions - *Undeclared*
    - 3.11: Chapter 3 References - *Undeclared*
  - 4: The Normal Distribution - *CC BY 4.0*
    - 4.1: Introduction - *CC BY 4.0*
    - 4.2: The Standard Normal Distribution - *CC BY 4.0*
    - 4.3: Using the Normal Distribution - *CC BY 4.0*
    - 4.4: Chapter 4 Key Terms - *CC BY 4.0*
    - 4.5: Chapter 4 Review - *CC BY 4.0*
    - 4.6: Chapter 4 Formula Review - *Undeclared*
    - 4.7: Chapter 4 Homework - *Undeclared*
    - 4.8: Chapter 4 Solutions - *Undeclared*
    - 4.9: Chapter 4 References - *Undeclared*
  - 5: The Central Limit Theorem - *CC BY 4.0*
    - 5.1: Introduction to the Central Limit Theorem - *CC BY 4.0*
    - 5.2: The Central Limit Theorem for Sample Means - *CC BY 4.0*
    - 5.3: Using the Central Limit Theorem - *CC BY 4.0*
    - 5.4: Chapter 5 Key Terms - *CC BY 4.0*
    - 5.5: Chapter 5 Review - *Undeclared*
    - 5.6: Chapter 5 Formula Review - *Undeclared*
    - 5.7: Chapter 5 Homework - *Undeclared*
    - 5.8: Chapter 5 Solutions - *Undeclared*

- 5.9: Chapter 5 References - *Undeclared*
- 6: Confidence Intervals - *CC BY 4.0*
  - 6.1: Introduction - *CC BY 4.0*
  - 6.2: A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size - *CC BY 4.0*
  - 6.3: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case - *CC BY 4.0*
  - 6.4: A Confidence Interval for A Population Proportion - *CC BY 4.0*
  - 6.5: Chapter 6 Key Terms - *CC BY 4.0*
  - 6.6: Chapter 6 Review - *Undeclared*
  - 6.7: Chapter 6 Formula Review - *Undeclared*
  - 6.8: Chapter 6 Homework - *Undeclared*
  - 6.9: Chapter 6 Solutions - *Undeclared*
  - 6.10: Chapter 6 References - *Undeclared*
- 7: Hypothesis Testing with One Sample - *CC BY 4.0*
  - 7.1: Introduction to Hypothesis Testing - *CC BY 4.0*
  - 7.2: Null and Alternative Hypotheses - *CC BY 4.0*
  - 7.3: Outcomes and Type I and Type II Errors - *CC BY 4.0*
  - 7.4: Distribution Needed for Hypothesis Testing - *CC BY 4.0*
  - 7.5: Full Hypothesis Test Examples - *CC BY 4.0*
  - 7.6: Chapter 7 Key Terms - *CC BY 4.0*
  - 7.7: Chapter 7 Review - *Undeclared*
  - 7.8: Chapter 7 Formula Review - *Undeclared*
  - 7.9: Chapter 7 Homework - *Undeclared*
  - 7.10: Chapter 7 Solutions - *Undeclared*
  - 7.11: Chapter 7 References - *Undeclared*
- 8: Hypothesis Testing with Two Samples - *CC BY 4.0*
  - 8.1: Introduction - *CC BY 4.0*
  - 8.2: Comparing Two Independent Population Means - *CC BY 4.0*
  - 8.3: Cohen's Standards for Small, Medium, and Large Effect Sizes - *CC BY 4.0*
  - 8.4: Comparing Two Independent Population Proportions - *CC BY 4.0*
  - 8.5: Matched or Paired Samples - *Undeclared*
  - 8.6: Chapter 8 Key Terms - *Undeclared*
  - 8.7: Chapter 8 Review - *Undeclared*
  - 8.8: Chapter 8 Formula Review - *Undeclared*
  - 8.9: Chapter 8 Homework - *Undeclared*
  - 8.10: Chapter 8 Solutions - *Undeclared*
  - 8.11: Chapter 8 References - *Undeclared*
- 9: F-Distribution and One-Way ANOVA - *CC BY 4.0*
  - 9.1: Introduction - *CC BY 4.0*
  - 9.2: One-Way ANOVA - *CC BY 4.0*
  - 9.3: The F-Distribution and the F-Ratio - *CC BY 4.0*
  - 9.4: Chapter 9 Key Terms - *CC BY 4.0*
  - 9.5: Chapter 9 Review - *Undeclared*
  - 9.6: Chapter 9 Formula Review - *Undeclared*
  - 9.7: Chapter 9 Homework - *Undeclared*
  - 9.8: Chapter 9 Solutions - *Undeclared*
  - 9.9: Chapter 9 References - *Undeclared*
- 10: Linear Regression and Correlation - *CC BY 4.0*
  - 10.1: Introduction - *CC BY 4.0*
  - 10.2: The Correlation Coefficient r - *CC BY 4.0*
  - 10.3: Testing the Significance of the Correlation Coefficient - *CC BY 4.0*
  - 10.4: Linear Equations - *CC BY 4.0*
  - 10.5: The Regression Equation - *CC BY 4.0*
  - 10.6: How to Use Microsoft Excel® for Regression Analysis - *Undeclared*
  - 10.7: Chapter 10 Key Terms - *Undeclared*
  - 10.8: Chapter 10 Review - *Undeclared*
  - 10.9: Chapter 10 Homework - *Undeclared*
  - 10.10: Chapter 10 Solutions - *Undeclared*
- 11: Appendices - *CC BY 4.0*
  - 11.1: A - Statistical Tables - *CC BY 4.0*
  - 11.2: B - Mathematical Phrases, Symbols, and Formulas - *CC BY 4.0*
  - 11.3: C - Reporting Statistics in APA Style - *Undeclared*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*