MATH 11: ELEMENTARY STATISTICS



Jim Yang
Fresno City College



Math 11 Elementary Statistics

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

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

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



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

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

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

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

This text was compiled on 04/19/2026

TABLE OF CONTENTS

Licensing

1: Introduction to Statistics

- 1.1: Definitions of Statistics and Key Terms
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 1.3: Frequency, Frequency Tables, and Levels of Measurement
  - 1.3.1: Experimental Design and Ethics

2: Data Displays

- 2.1: Data, Sampling, and Variation in Data and Sampling
- 2.2: Histogram
- 2.3: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

3: Descriptive Statistics

- 3.1: Measures of Center
- 3.2: Measures of Variability
- 3.3: Relative Position of Data
- 3.4: The Empirical Rule and Chebyshev's Theorem

4: Probability Topics

- 4.1: Introduction
  - 4.1.1: Terminology
  - 4.1.2: Independent and Mutually Exclusive Events
- 4.2: Addition and Multiplication Rule of Probability
- 4.3: Conditional Probability using Contingency Tables
- 4.E: Probability Topics (Exericses)

5: Discrete Random Variables

- 5.1: Random Variables
  - 5.1.1: Probability Distributions for Discrete Random Variables
- 5.2: The Binomial Distribution
- 5.E: Discrete Random Variables (Exercises)

6: Continuous Random Variables

- 6.1: The Standard Normal Distribution
  - 6.1.1: Continuous Random Variables
  - 6.1.2: The Standard Normal Distribution
- 6.2: The General Normal Distribution
- 6.3: The Central Limit Theorem for Sample Means
  - 6.3E: The Central Limit Theorem for Sample Means (Exercises)

7: Estimation

- 7.1: Estimation of a Population Proportion
- 7.2: Estimation of a Population Mean
  - 7.2.1: Large Sample Estimation of a Population Mean
  - 7.2.2: Small Sample Estimation of a Population Mean
- 7.3: Sample Size Considerations
- 7.E: Estimation (Exercises)

8: Testing Hypotheses

- 8.1: The Elements of Hypothesis Testing
- 8.2: Tests for a Population Mean
  - 8.2.1: The Observed Significance of a Test
  - 8.2.2: Small Sample Tests for a Population Mean
- 8.3: Tests for a Population Proportion
- 8.E: Testing Hypotheses (Exercises)

9: Two-Sample Problems

- 9.1: Two Population Proportions
- 9.2: Two Population Means - Independent Samples
  - 9.2.1: Large, Independent Samples
  - 9.2.2: Small, Independent Samples
- 9.3: Two Population Means - Paired Samples
- 9.4: Sample Size Considerations
- 9.E: Two-Sample Problems (Exercises)

10: Linear Regression and Correlation

- 10.1: Introduction to Linear Regression and Correlation
  - 10.1.1: Linear Equations
    - 10.1.1E: Linear Equations (Exercises)
  - 10.1.2: Scatter Plots
    - 10.1.2E: Scatter Plots (Exercises)
- 10.2: The Regression Equation and Correlation Coefficient
  - 10.2E: The Regression Equation (Exercise)
- 10.3: Testing for Significance Linear Correlation
  - 10.3E: Testing the Significance of the Correlation Coefficient (Exercises)
- 10.4: Prediction
  - 10.4E: Prediction (Exercises)
- 10.E: Linear Regression and Correlation (Exercises)

11: Chi-Square Tests

- 11.1: Chi-Square Tests for Independence
- 11.2: Chi-Square One-Sample Goodness-of-Fit Tests

12: Analysis of Variance

- 12.1: F-Tests
- 12.2: F-Tests in One-Way ANOVA

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

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

CHAPTER OVERVIEW

1: Introduction to Statistics

In this chapter we will introduce some basic terminology and lay the groundwork for the course. We will explain in general terms what statistics and probability are and the problems that these two areas of study are designed to solve.

1.1: Definitions of Statistics and Key Terms

1.2: Data, Sampling, and Variation in Data and Sampling

1.3: Frequency, Frequency Tables, and Levels of Measurement

1.3.1: Experimental Design and Ethics

This page titled 1: Introduction to Statistics is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

1.1: Definitions of Statistics 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.

Collaborative Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5; 5.5; 6; 6; 6; 6.5; 6.5; 6.5; 6.5; 7; 7; 8; 8; 9

The dot plot for this data would be as follows:

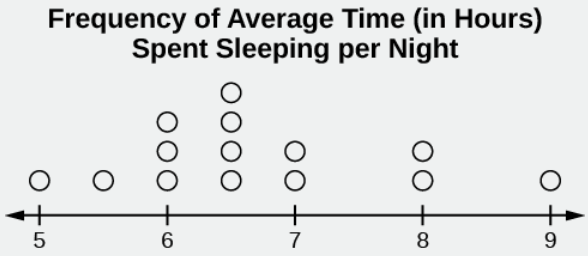


Figure **1.1.1**

- Does your dot plot look the same as or different from the example? Why?
- If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?
- Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

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.

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. A **parameter** is a number that is a property of the population. 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**, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in 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).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

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.

> 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 1.1.1

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 survey 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 **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 1.1.1

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.

Answer

- The **population** is all families with children attending Knoll Academy.
- The **sample** is a random selection of 100 families with children attending Knoll Academy.
- The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.
- The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.
- The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.
- The **data** are the dollar amounts spent by the families. Examples of the data are $65, $75, and $95.

Example 1.1.2

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 1.1.3

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:

Speed at which Cars Crashed	Location of "drive" (i.e. dummies)
35 miles/hour	Front Seat

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 1.1.4

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.

Collaborative Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

1. The Data and Story Library, https://dasl.datadescription.com/ (accessed May 1, 2013).

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:

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

Researcher B:

3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the key terms refer to in the example for Researcher A.

Exercise 1.1.2

population

Answer

AIDS patients.

Exercise 1.1.3

sample

Exercise 1.1.4

parameter

Answer

The average length of time (in months) AIDS patients live after treatment.

Exercise 1.1.5

statistic

Exercise 1.1.6

variable

Answer

$X =$ the length of time (in months) AIDS patients live after treatment

Glossary

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.

Average

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

Categorical Variable

variables that take on values that are names or labels

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)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

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

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 successes divided by the total number in the sample

Representative Sample

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

Sample

a subset of the population studied

Statistic

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

Variable

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

---

This page titled 1.1: Definitions of Statistics and Key Terms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.2: Data, Sampling, and Variation in Data and Sampling

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

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative 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 data over qualitative 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.

Quantitative 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.

All data that are the result of measuring are *quantitative continuous data* assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Sample of Quantitative Discrete Data

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 1.2.1

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?

Answer

quantitative discrete data

Sample of Quantitative Continuous Data

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** because weights are measured.

Exercise 1.2.2

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?

Answer

quantitative continuous data

Exercise 1.2.3

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

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

Solution

One Possible Solution:

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

Try to identify additional data sets in this example.

Sample of qualitative data

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 data**.

Exercise 1.2.4

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?

Answer

qualitative data

Collaborative Exercise 1.2.1

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. where you go on vacation
d. the distance it is from your home to the nearest grocery store
e. the number of classes you take per school year.
f. the tuition for your classes
g. the type of calculator you use
h. movie ratings
i. political party preferences
j. weights of sumo wrestlers
k. amount of money (in dollars) won playing poker
l. number of correct answers on a quiz
m. peoples' attitudes toward the government
n. IQ scores (This may cause some discussion.)

Answer

Items a, e, f, k, and l are quantitative discrete; items d, j, and n are quantitative continuous; items b, c, g, h, i, and m are qualitative.

Exercise 1.2.5

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.

Answer

quantitative discrete

Exercise 1.2.6

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 pie chart Figure 1.2.1. What type of data does this graph show?

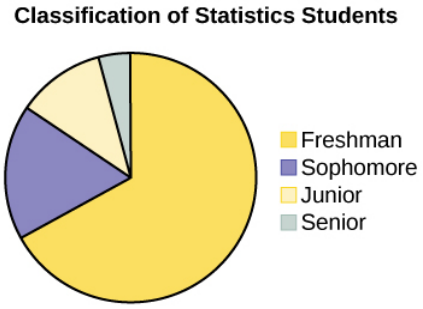


Figure 1.2.1

Answer

This pie chart shows the students in each year, which is **qualitative data**.

Exercise 1.2.7

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.

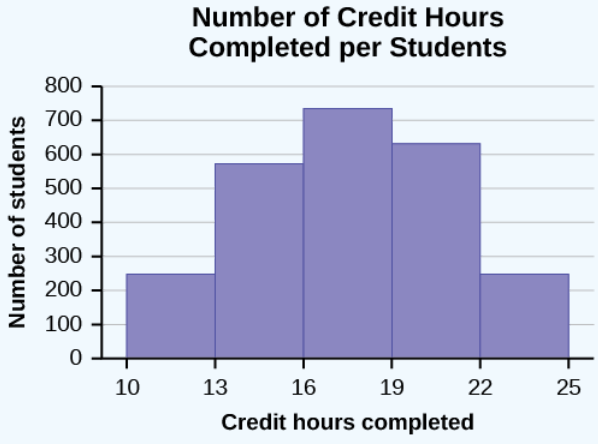


Figure 1.2.2

What type of data does this graph show?

Answer

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

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 **1.2.1**: Fall Term 2007 (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. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

- In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.
- In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.
- A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figures **1.2.3** and **1.2.4** and determine which graph (pie or bar) you think displays the comparisons better.

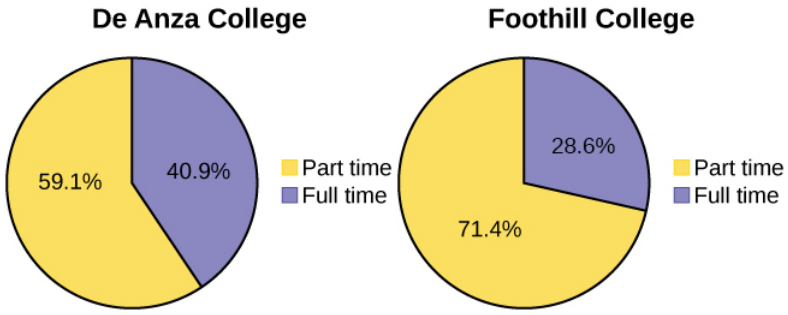


Figure **1.2.3**: Pie Charts

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context. Our choice also depends on what we are using the data for.

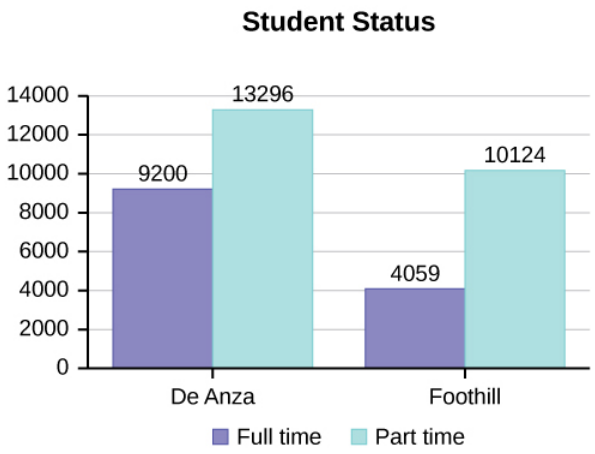


Figure 1.2.4: Bar chart

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

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

Table 1.2.2: 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%

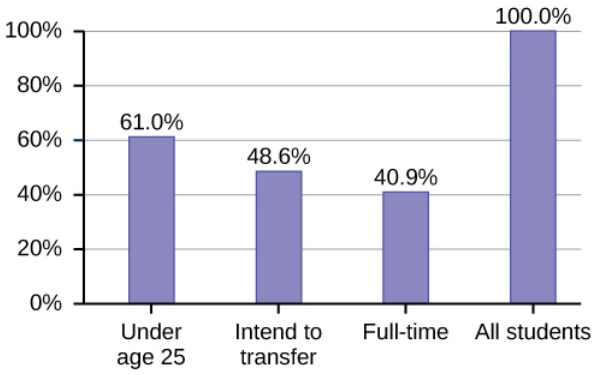


Figure 1.2.2: Bar chart of data in Table 1.2.2.

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. In this situation, create a bar graph and not a pie chart.

Table 1.2.2: 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%

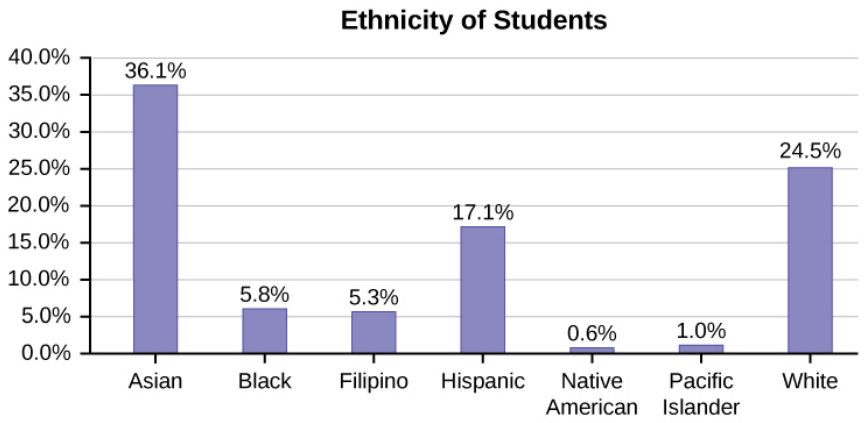


Figure **1.2.3**: Enrollment of De Anza College (Spring 2010)

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 **1.2.4** can be difficult to understand visually. The graph in Figure **1.2.5** is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

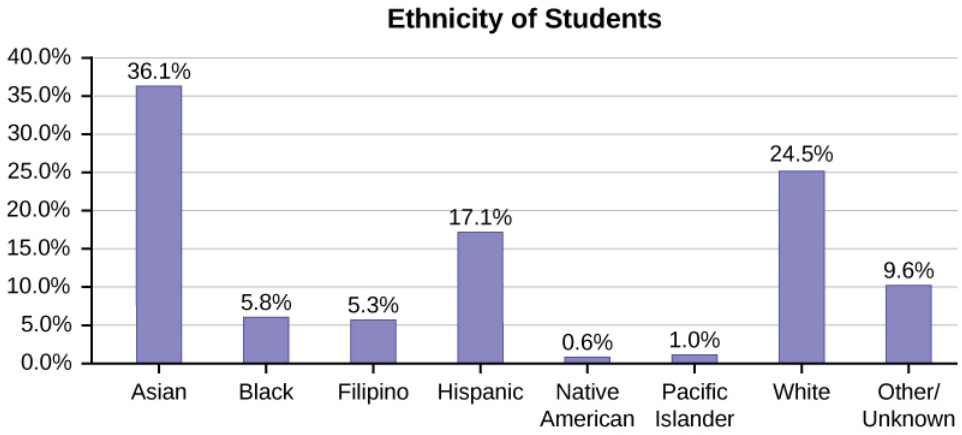


Figure **1.2.4**: Bar Graph with Other/Unknown Category

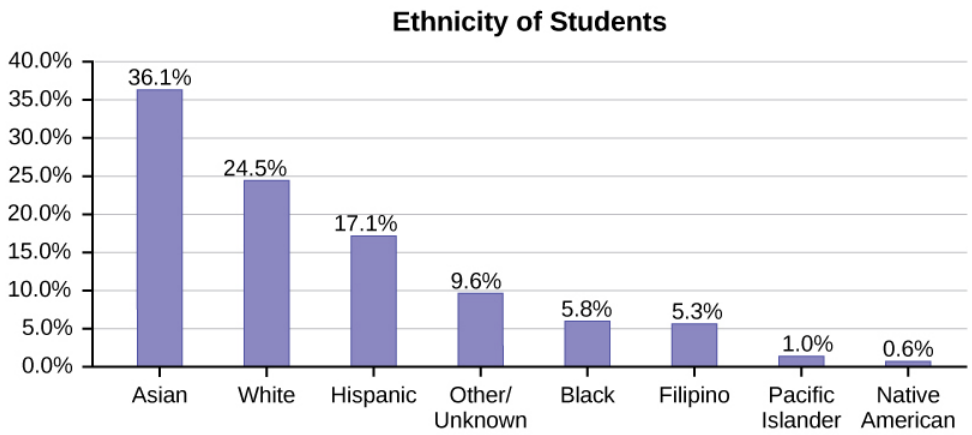


Figure 1.2.5: Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Figure 1.2.6 is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 1.2.6.

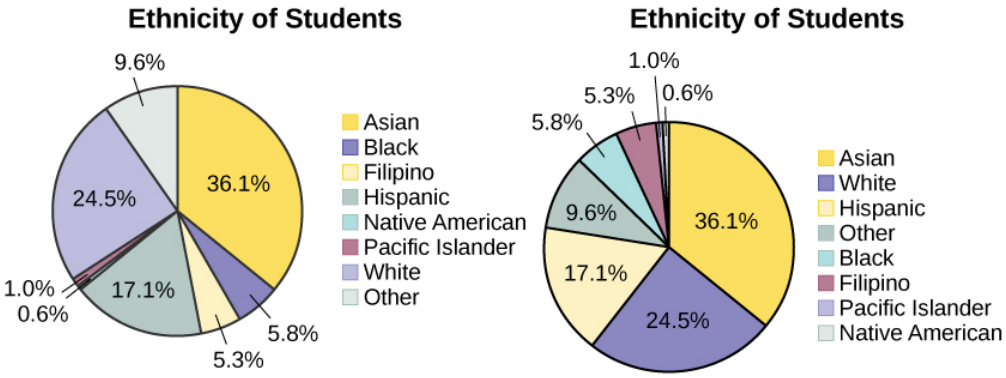


Figure 1.2.6.

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 by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in Table 1.2.2:

Table 1.2.3: Class Roster

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

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

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

To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.

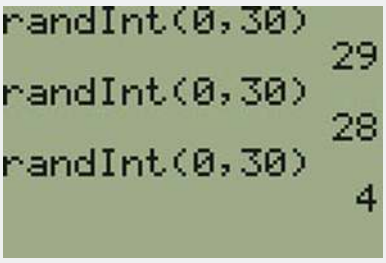


Figure 1.2.7

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 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. 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.

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 n^{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 choose every fiftieth name thereafter 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 simple 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, **a 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.

Exercise 1.2.8

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 1.2.9: Calculator

You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

a. Create a stratified sample by column. Pick three quiz scores randomly from each column.
- Number each row one through ten.
- On your calculator, press Math and arrow over to PRB.
- For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
- Repeat for columns two through six.
- These 18 quiz scores are a stratified sample.

b. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
- Press MATH and arrow over to PRB.
- Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
- The two numbers are for two of the columns.
- The quiz scores (20 of them) in these 2 columns are the cluster sample.

c. Create a simple random sample of 15 quiz scores.
- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER 15 times and record the numbers.
- Record the quiz scores that correspond to these numbers.
- These 15 quiz scores are the systematic sample.

d. Create a systematic sample of 12 quiz scores.
- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example 1.2.10

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

Exercise 1.2.11

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

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Answer

stratified

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 1.2.12: Sampling

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

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 parttime 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

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 1.2.12

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?

Answer

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

Collaborative Exercise 1.2.8

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
b. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
c. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

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) 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. You will learn why when you study confidence intervals.

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

Collaborative Exercise 1.2.8

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in the following table ("frequency" is the number of times a particular face of the die occurs):

First Experiment (20 rolls)		Second Experiment (20 rolls)	
Face on Die	**Frequency**	**Face on Die**	**Frequency**
1			
2			
3			
4			
5			
6			

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

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.

References

1. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
2. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
3. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
4. Data from www.bookofodds.com/Relationsh...-the-President
5. 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).
6. "The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).
7. "Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/ga...9362004.aspx#4 (accessed May 1, 2013).
8. The Data and Story Library, lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).
9. LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/f...hts.html#focus (accessed May 1, 2013).
10. Data from San Jose Mercury News

Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, 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.

Footnotes

1. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
2. Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
3. Frequently Asked Questions, Pew Research Center for the People & the Press, www.people-press.org/methodol...wer-your-polls (accessed May 1, 2013).

Glossary

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.

Convenience Sampling

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

Discrete Random Variable

a random variable (RV) whose outcomes are counted

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.

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Sampling

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

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 labels. These randomly selected labels identify the members of your sample.

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.

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.

This page titled 1.2: Data, Sampling, and Variation in Data and Sampling is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.3: Frequency, Frequency Tables, and 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 possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

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**. 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. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. 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 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 lists the different data values in ascending order and their frequencies.

Table **1.3.1**: Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Definition: Relative Frequency

A frequency is the number of times a value of the data occurs. According to Table Table **1.3.1**, 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.

Definition: Relative frequencies

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

Table **1.3.2**: Frequency Table of Student Work Hours with Relative Frequencies

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

The sum of the values in the relative frequency column of Table **1.3.2** is $\frac{20}{20}$, or 1.

Definition: Cumulative Relative Frequency

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 **1.3.3**.

Table **1.3.3**: Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

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

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

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 **1.3.4** represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table **1.3.4**: Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 = 1.00$
	Total = 100	**Total = 1.00**	

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

- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches

- 73.95 to 75.95 inches

This example is used again in Descriptive Statistics, where the method used to compute the intervals will be explained.

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

Exercise 1.3.1

a. From the Table **1.3.4**, find the percentage of heights that are less than 65.95 inches.
b. Find the percentage of heights that fall between 61.95 and 65.95 inches.

Answer

a. If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are $5 + 3 + 15 = 23$ players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.
b. Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$ or 18%.

Exercise 1.3.2

Table **1.3.5** shows the amount, in inches, of annual rainfall in a sample of towns.

a. Find the percentage of rainfall that is less than 9.01 inches.
b. Find the percentage of rainfall that is between 6.99 and 13.05 inches.

Table **1.3.5**

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	

Answer

a. 0.56 or 56
b. $0.30 + 0.16 + 0.18 = 0.64$ or 64

Exercise 1.3.3

Use the heights of the 100 male semiprofessional soccer players in Table **1.3.4**. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.

c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
e. What kind of data are the heights?
f. 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

a. 29%
b. 36%
c. 77%
d. 87
e. quantitative continuous
f. get rosters from each team and choose a simple random sample from each

Exercise 1.3.4

From Table **1.3.5**, find the number of towns that have rainfall between 2.95 and 9.01 inches.

Answer

$6 + 7 + 15 = 28$ towns

Collaborative Exercise 1.3.7

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

a. What percentage of the students in your class have no siblings?
b. What percentage of the students have from one to three siblings?
c. What percentage of the students have fewer than three siblings?

Example 1.3.7

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 **1.3.6** was produced:

Table **1.3.6**: 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	$\frac{3}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Answer

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
b. 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.
c. $\frac{5}{19}$
d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Exercise 1.3.8

Table **1.3.5** 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?

Answer

$\frac{9}{50}$

Example 1.3.9

Table **1.3.7** contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Table **1.3.7**: Total Number of Deaths Worldwide as a Result of Earthquakes

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	823,356

Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?
f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Answer

a. 97,118 (11.8%)
b. 41.6%
c. 67,092/823,356 or 0.081 or 8.1 %
d. 27.8%
e. Quantitative discrete
f. Quantitative continuous

? Exercise 1.3.10

Table **1.3.8** contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Table **1.3.8**:

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		

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.

Answer

a. 190,800 (29.2%)
b. 24.9%
c. 260,086/653,782 or 39.8%
d. 4.6%
e. 75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

References

1. “State & County QuickFacts,” U.S. Census Bureau. quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).
2. “State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau. quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).
3. “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).
4. “Levels of Measurement,” infinity.cos.edu/faculty/wood...ata_Levels.htm (accessed May 1, 2013).
5. Courtney Taylor, “Levels of Measurement,” about.com, http://statistics.about.com/od/Helpa...easurement.htm (accessed May 1, 2013).
6. David Lane. “Levels of Measurement,” Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

Review

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 starting point; the differences can be measured, but there is no such thing as a ratio.
- **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.

Exercise 1.3.11

What type of measure scale is being used? Nominal, ordinal, interval or ratio.

a. High school soccer players classified by their athletic ability: Superior, Average, Above average
b. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
c. The colors of crayons in a 24-crayon box
d. Social security numbers
e. Incomes measured in dollars
f. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
g. Political outlook: extreme left, left-of-center, right-of-center, extreme right
h. Time of day on an analog watch
i. The distance in miles to the closest grocery store
j. The dates 1066, 1492, 1644, 1947, and 1944
k. The heights of 21–65 year-old women
l. Common letter grades: A, B, C, D, and F

Answer

a. ordinal
b. interval
c. nominal
d. nominal
e. ratio
f. ordinal
g. nominal
h. interval
i. ratio
j. interval
k. ratio
l. ordinal

Glossary

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.

Frequency

the number of times a value of the data occurs

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 to the total number of outcomes

This page titled 1.3: Frequency, Frequency Tables, and Levels of Measurement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.3.1: 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 **explanatory variable**. The affected variable is called the response variable. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable 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 prove 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 prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove 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 prove 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.*[1]

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 1.3.1.1

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Answer

- The population is men aged 50 to 84

- The *population* is men aged 30 to 84.
- The *sample* is the 400 men who participated.
- The *experimental units* are the individual men in the study.
- The *explanatory variable* is oral medication.
- The *treatments* are aspirin and a placebo.
- The *response variable* is whether a subject had a heart attack.

Example 1.3.1.2

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

a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
b. There are two treatments: a floral-scented mask and an unscented mask.
c. 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.
d. 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.

Example 1.3.1.3

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Answer

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

Exercise 1.3.1.4

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

a. Describe the explanatory and response variables in the study.
b. What are the treatments?
c. What should you consider when selecting participants?
d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
e. Identify any lurking variables that could interfere with this study.
f. How can blinding be used in this study?

Answer

a. Explanatory: presence of distraction from texting; response: response time measured in seconds

b. Driving without distraction and driving while texting
c. Answers will vary. Possible responses: Do participants regularly send and receive text messages? How long has the subject been driving? What is the age of the participants? Do participants have similar texting and driving experience?
d. This is not a good plan because it compares drivers with different abilities. It would be better to assign both treatments to each participant in random order.
e. Possible responses include: texting ability, driving experience, type of phone.
f. The researchers observing the trials and recording response time could be blinded to the treatment being applied.

Ethics

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

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including *Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology, British Journal of Social Psychology,* and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.[2]

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

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

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[3] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

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

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

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

- Risks to participants must be minimized and reasonable with respect to projected benefits.

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

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

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

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

Example 1.3.1.5

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

A researcher is collecting data in a community.

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

Answer

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

Exercise 1.3.1.6

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

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

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

Answer

a. This is not necessarily a problem. The study should be monitored carefully, however, to ensure that the company is not pressuring researchers to return biased results.
b. If the researchers truly want to determine the favorite brand of juice, then researchers should ask teens to compare different brands of the same type of juice. Choosing a sweet juice to compare against a sharp-flavored juice will not lead to an accurate comparison of brand quality.
c. Participants could be biased by the knowledge. The results may be different from those obtained in a blind taste test.
d. The commercial tells the truth, but not the whole truth. It leads consumers to believe that Brand X was preferred by more participants than Brand Y while the opposite is true.

References

1. "Vitamin E and Health," Nutrition Source, Harvard School of Public Health, www.hsph.harvard.edu/nutritio...rce/vitamin-e/ (accessed May 1, 2013).
2. Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).
3. 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).
4. The Data and Story Library, lib.stat.cmu.edu/DASL/Stories...dLearning.html (accessed May 1, 2013).
5. 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).
6. "Earthquake Information by Year," U.S. Geological Survey. earthquake.usgs.gov/earthquak...archives/year/ (accessed May 1, 2013).
7. "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).
8. Data from www.businessweek.com (accessed May 1, 2013).
9. Data from www.forbes.com (accessed May 1, 2013).
10. "America's Best Small Companies," http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).
11. 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.
12. "April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), www.dot.gov/airconsumer/april...onsumer-report (accessed May 1, 2013).
13. Lori Alden, "Statistics can be Misleading," econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).
14. 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).

Review

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.

"An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule."[4] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

? Exercise 1.3.1.7

Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

? Exercise 1.3.1.7

Discuss potential violations of the rule requiring informed consent.

a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
b. A research study is designed to investigate a new children's allergy medication.
c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Answer

a. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

Footnotes

[1] McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

[2] Y.udhijit Bhattacharjee, "The Mind of a Con Man," Magazine, New York Times, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/ma...src=dayp&_r=2& (accessed May 1, 2013).

[3] "Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," Tillburg University, November 28, 2012, www.tilburguniversity.edu/upl...012_UK_web.pdf (accessed May 1, 2013).

[4] Andrew Gelman, "Open Data and Open Methods," Ethics and Statistics, http://www.stat.columbia.edu/~gelman...nceEthics1.pdf (accessed May 1, 2013).

Glossary

Explanatory Variable
the independent variable in an experiment; the value controlled by researchers

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

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

Experimental Unit
any individual or object to be measured

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

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

Control Group

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

Informed Consent

Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board

a committee tasked with oversight of research programs that involve human subjects

Placebo

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

Blinding

not telling participants which treatment a subject is receiving

Double-blinding

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

---

This page titled 1.3.1: Experimental Design and Ethics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

2: Data Displays

2.1: Data, Sampling, and Variation in Data and Sampling

2.2: Histogram

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

---

2: Data Displays is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.1: Data, Sampling, and Variation in Data and Sampling

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

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative 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 data over qualitative 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.

Quantitative 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.

All data that are the result of measuring are *quantitative continuous data* assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Sample of Quantitative Discrete Data

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 2.1.1

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?

Answer

quantitative discrete data

Sample of Quantitative Continuous Data

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** because weights are measured.

Exercise 2.1.2

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?

Answer

quantitative continuous data

Exercise 2.1.3

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

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

Solution

One Possible Solution:

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

Try to identify additional data sets in this example.

Sample of qualitative data

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 data**.

Exercise 2.1.4

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?

Answer

qualitative data

Collaborative Exercise 2.1.1

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. where you go on vacation
d. the distance it is from your home to the nearest grocery store
e. the number of classes you take per school year.
f. the tuition for your classes
g. the type of calculator you use
h. movie ratings
i. political party preferences
j. weights of sumo wrestlers
k. amount of money (in dollars) won playing poker
l. number of correct answers on a quiz
m. peoples' attitudes toward the government
n. IQ scores (This may cause some discussion.)

Answer

Items a, e, f, k, and l are quantitative discrete; items d, j, and n are quantitative continuous; items b, c, g, h, i, and m are qualitative.

Exercise 2.1.5

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.

Answer

quantitative discrete

Exercise 2.1.6

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 pie chart Figure 2.1.1. What type of data does this graph show?

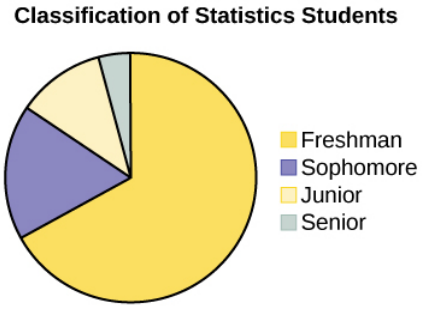


Figure 2.1.1

Answer

This pie chart shows the students in each year, which is **qualitative data**.

Exercise 2.1.7

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.

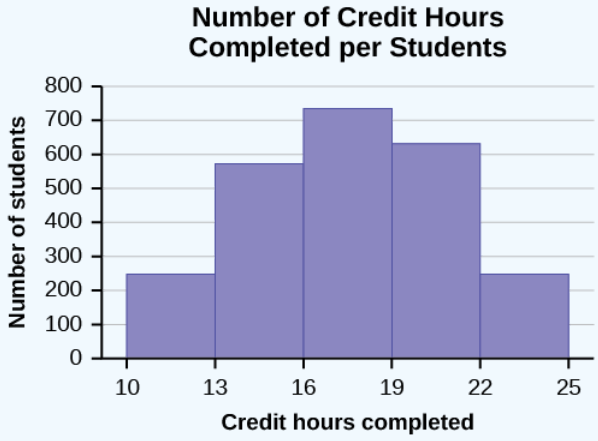


Figure 2.1.2

What type of data does this graph show?

Answer

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

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 **2.1.1**: Fall Term 2007 (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. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

- In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.
- In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.
- A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figures **2.1.3** and **2.1.4** and determine which graph (pie or bar) you think displays the comparisons better.

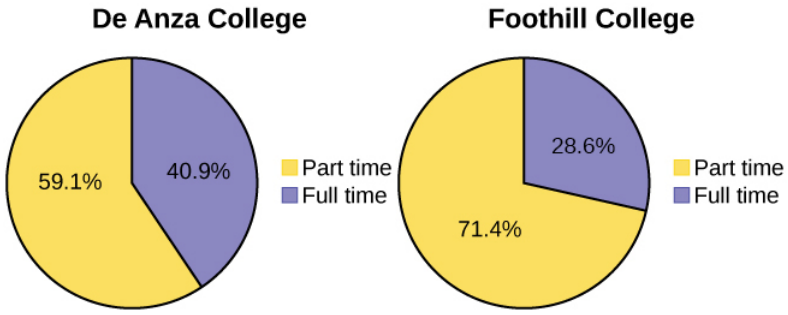


Figure **2.1.3**: Pie Charts

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context. Our choice also depends on what we are using the data for.

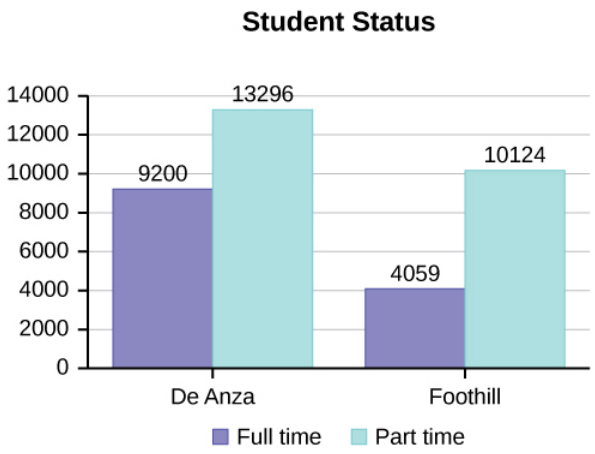


Figure **2.1.4**: Bar chart

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

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

Table **2.1.2**: 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%

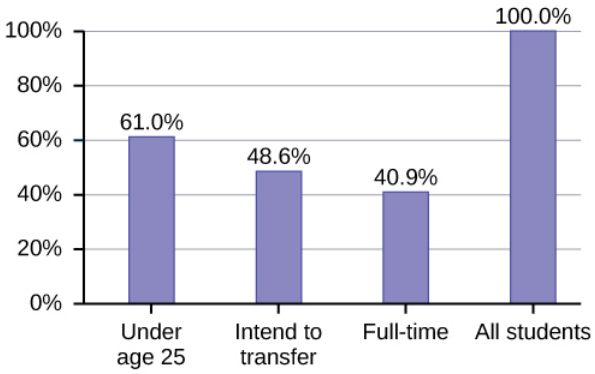


Figure **2.1.2**: Bar chart of data in Table **2.1.2**.

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. In this situation, create a bar graph and not a pie chart.

Table **2.1.2**: 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%

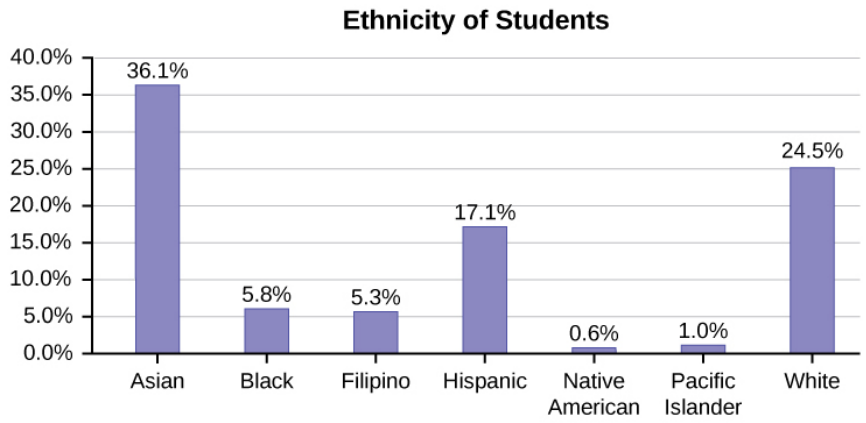


Figure **2.1.3**: Enrollment of De Anza College (Spring 2010)

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 **2.1.4** can be difficult to understand visually. The graph in Figure **2.1.5** is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

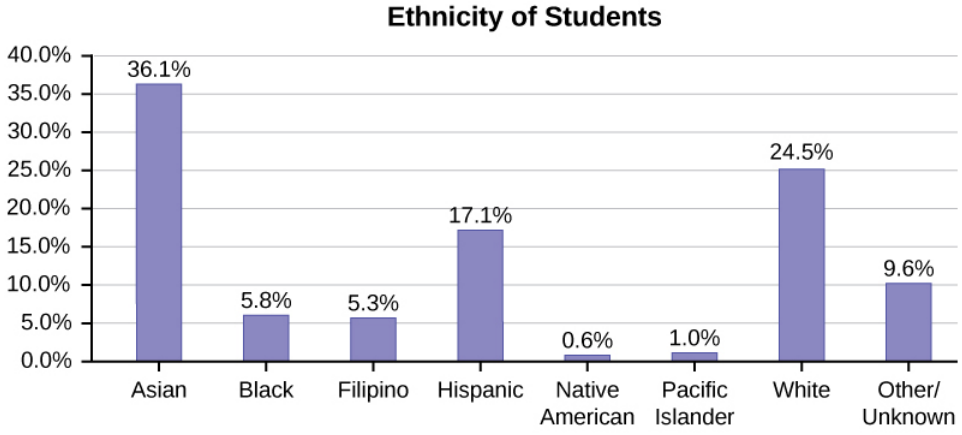


Figure **2.1.4**: Bar Graph with Other/Unknown Category

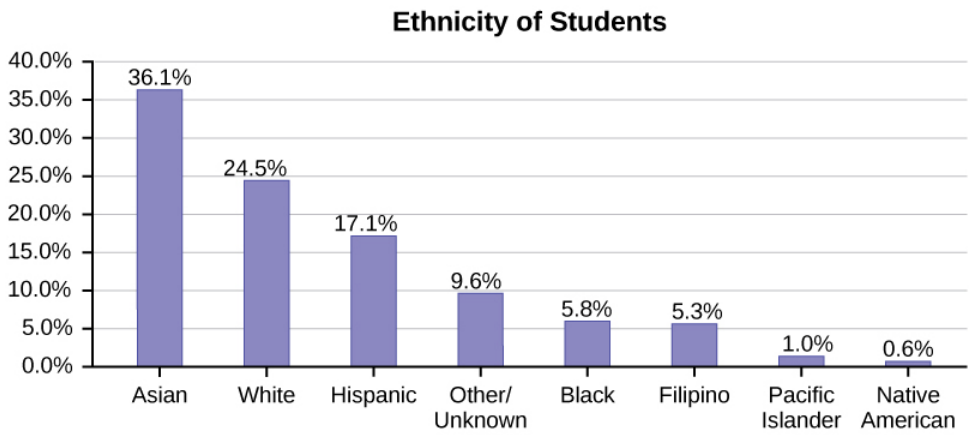


Figure 2.1.5: Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Figure 2.1.6 is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 2.1.6.

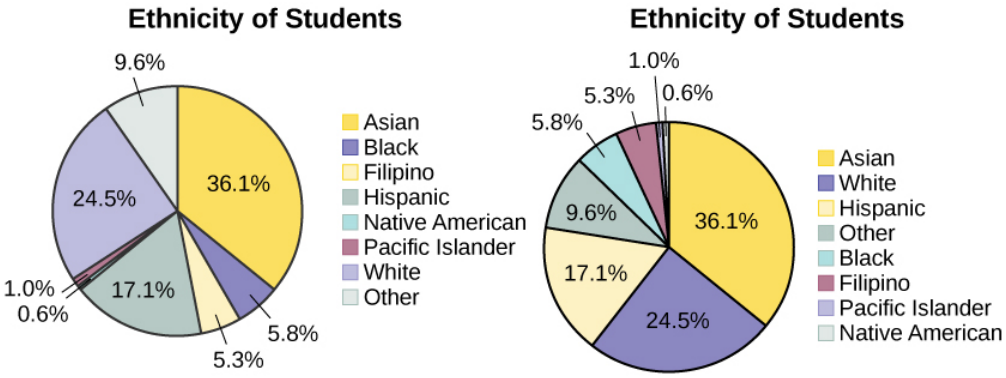


Figure 2.1.6.

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 by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in Table 2.1.2:

Table 2.1.3: Class Roster

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

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

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

To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.

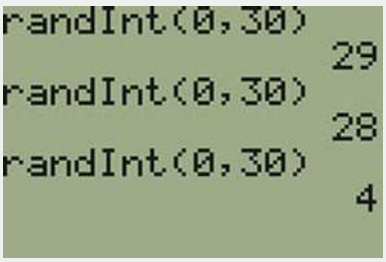


Figure 2.1.7

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 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. 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.

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 n^{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 choose every fiftieth name thereafter 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 simple 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, **a 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.

Exercise 2.1.8

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 2.1.9: Calculator

You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

a. Create a stratified sample by column. Pick three quiz scores randomly from each column.
- Number each row one through ten.
- On your calculator, press Math and arrow over to PRB.
- For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
- Repeat for columns two through six.
- These 18 quiz scores are a stratified sample.

b. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
- Press MATH and arrow over to PRB.
- Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
- The two numbers are for two of the columns.
- The quiz scores (20 of them) in these 2 columns are the cluster sample.

c. Create a simple random sample of 15 quiz scores.
- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER 15 times and record the numbers.
- Record the quiz scores that correspond to these numbers.
- These 15 quiz scores are the systematic sample.

d. Create a systematic sample of 12 quiz scores.
- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example 2.1.10

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

Exercise 2.1.11

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

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Answer

stratified

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 2.1.12: Sampling

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

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 parttime 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

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 2.1.12

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?

Answer

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

Collaborative Exercise 2.1.8

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
b. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
c. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

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) 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. You will learn why when you study confidence intervals.

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

Collaborative Exercise 2.1.8

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in the following table ("frequency" is the number of times a particular face of the die occurs):

First Experiment (20 rolls)		Second Experiment (20 rolls)	
Face on Die	**Frequency**	**Face on Die**	**Frequency**
1			
2			
3			
4			
5			
6			

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

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.

References

1. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
2. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
3. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
4. Data from www.bookofodds.com/Relationsh...-the-President
5. 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).
6. "The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).
7. "Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/ga...9362004.aspx#4 (accessed May 1, 2013).
8. The Data and Story Library, lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).
9. LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/f...hts.html#focus (accessed May 1, 2013).
10. Data from San Jose Mercury News

Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, 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.

Footnotes

1. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
2. Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
3. Frequently Asked Questions, Pew Research Center for the People & the Press, www.people-press.org/methodol...wer-your-polls (accessed May 1, 2013).

Glossary

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.

Convenience Sampling

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

Discrete Random Variable

a random variable (RV) whose outcomes are counted

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.

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Sampling

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

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 labels. These randomly selected labels identify the members of your sample.

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.

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.

This page titled 2.1: Data, Sampling, and Variation in Data and Sampling is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.2: Histogram

Learning Objectives

- To learn to interpret the meaning of three graphical representations of sets of data: stem and leaf diagrams, frequency histograms, and relative frequency histograms.

A well-known adage is that "a picture is worth a thousand words." This saying proves true when it comes to presenting statistical information in a data set. There are many effective ways to present data graphically. The three graphical tools that are introduced in this section are among the most commonly used and are relevant to the subsequent presentation of the material in this book.

Stem and Leaf Diagrams

Suppose 30 students in a statistics class took a test and made the following scores:

86	80	25	77	73	76	100	90	69	93
90	83	70	73	73	70	90	83	71	95
40	58	68	69	100	78	87	97	92	74

How did the class do on the test? A quick glance at the set of 30 numbers does not immediately give a clear answer. However the data set may be reorganized and rewritten to make relevant information more visible. One way to do so is to construct a stem and leaf diagram as shown in Figure 2.2.1 The numbers in the tens place, from 2 through 9, and additionally the number 10, are the "stems," and are arranged in numerical order from top to bottom to the left of a vertical line. The number in the units place in each measurement is a "leaf," and is placed in a row to the right of the corresponding stem, the number in the tens place of that measurement. Thus the three leaves $9, 8,$ and 9 in the row headed with the stem 6 correspond to the three exam scores in the $60s, 69$ (in the first row of data), 68 (in the third row), and 69 (also in the third row).

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

Figure 2.2.1: *Stem and Leaf Diagram*

The display is made even more useful for some purposes by rearranging the leaves in numerical order, as shown in Figure 2.2.2. Either way, with the data reorganized certain information of interest becomes apparent immediately. There are two perfect scores; three students made scores under 60; most students scored in the $70s, 80s$ and $90s$; and the overall average is probably in the high $70s$ or low $80s$.

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

Figure 2.2.2: *Ordered Stem and Leaf Diagram*

In this example the scores have a natural stem (the tens place) and leaf (the ones place). One could spread the diagram out by splitting each tens place number into lower and upper categories. For example, all the scores in the $80s$ may be represented on two

separate stems, lower $80s$ and upper $80s$:

8	0 3 3
8	6 7

The definitions of stems and leaves are flexible in practice. The general purpose of a stem and leaf diagram is to provide a quick display of how the data are distributed across the range of their values; some improvisation could be necessary to obtain a diagram that best meets that goal.

Note that all of the original data can be recovered from the stem and leaf diagram. This will not be true in the next two types of graphical displays.

Frequency Histograms

The stem and leaf diagram is not practical for large data sets, so we need a different, purely graphical way to represent data. A frequency histogram is such a device. We will illustrate it using the same data set from the previous subsection. For the 30 scores on the exam, it is natural to group the scores on the standard ten-point scale, and count the number of scores in each group. Thus there are two $100s$, seven scores in the $90s$, six in the $80s$, and so on. We then construct the diagram shown in Figure $2.2.3$ by drawing for each group, or class, a vertical bar whose length is the number of observations in that group. In our example, the bar labeled 100 is 2 units long, the bar labeled 90 is 7 units long, and so on. While the individual data values are lost, we know the number in each class. This number is called the frequency of the class, hence the name frequency histogram.

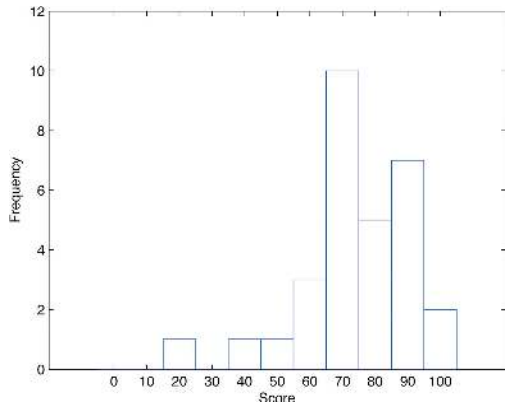


Figure $2.2.3$: *Frequency Histogram*

The same procedure can be applied to any collection of numerical data. Observations are grouped into several classes and the frequency (the number of observations) of each class is noted. These classes are arranged and indicated in order on the horizontal axis (called the x-axis), and for each group a vertical bar, whose length is the number of observations in that group, is drawn. The resulting display is a frequency histogram for the data. The similarity in Figure $2.2.1$ and Figure $2.2.3$ is apparent, particularly if you imagine turning the stem and leaf diagram on its side by rotating it a quarter turn counterclockwise.

Definition

In general, the definition of the classes in the frequency histogram is flexible. The general purpose of a frequency histogram is very much the same as that of a stem and leaf diagram, to provide a graphical display that gives a sense of data distribution across the range of values that appear.

We will not discuss the process of constructing a histogram from data since in actual practice it is done automatically with statistical software or even handheld calculators.

Relative Frequency Histograms

In our example of the exam scores in a statistics class, five students scored in the $80s$. The number 5 is the frequency of the group labeled "$80s$." Since there are 30 students in the entire statistics class, the proportion who scored in the $80s$ is $5/30$. The number $5/30$, which could also be expressed as $0.1\overline{6} \approx .1667$, or as 16.67%, is the relative frequency of the group labeled "$80s$." Every group

(the $70s$, the $80s$, and so on) has a relative frequency. We can thus construct a diagram by drawing for each group, or class, a vertical bar whose length is the relative frequency of that group. For example, the bar for the $80s$ will have length $5/30$ unit, not 5 units. The diagram is a relative frequency histogram for the data, and is shown in Figure 2.2.4. It is exactly the same as the frequency histogram except that the vertical axis in the relative frequency histogram is not frequency but relative frequency.

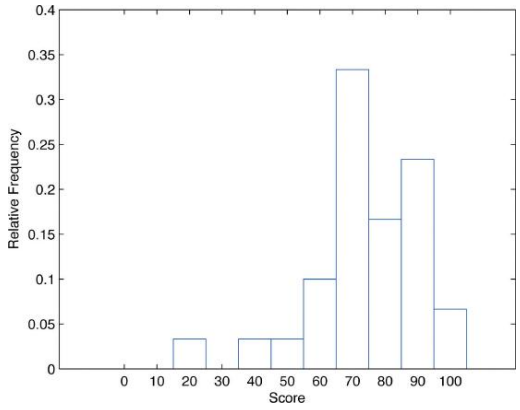


Figure 2.2.4: *Relative Frequency Histogram*

The same procedure can be applied to any collection of numerical data. Classes are selected, the relative frequency of each class is noted, the classes are arranged and indicated in order on the horizontal axis, and for each class a vertical bar, whose length is the relative frequency of the class, is drawn. The resulting display is a relative frequency histogram for the data. A key point is that now if each vertical bar has width 1 unit, then the total area of all the bars is 1 or 100%.

Although the histograms in Figure 2.2.3 and Figure 2.2.4 have the same appearance, the relative frequency histogram is more important for us, and it will be relative frequency histograms that will be used repeatedly to represent data in this text. To see why this is so, reflect on what it is that you are actually seeing in the diagrams that quickly and effectively communicates information to you about the data. It is the relative sizes of the bars. The bar labeled "$70s$" in either figure takes up $1/3$ of the total area of all the bars, and although we may not think of this consciously, we perceive the proportion $1/3$ in the figures, indicating that a third of the grades were in the $70s$. The relative frequency histogram is important because the labeling on the vertical axis reflects what is important visually: the relative sizes of the bars.

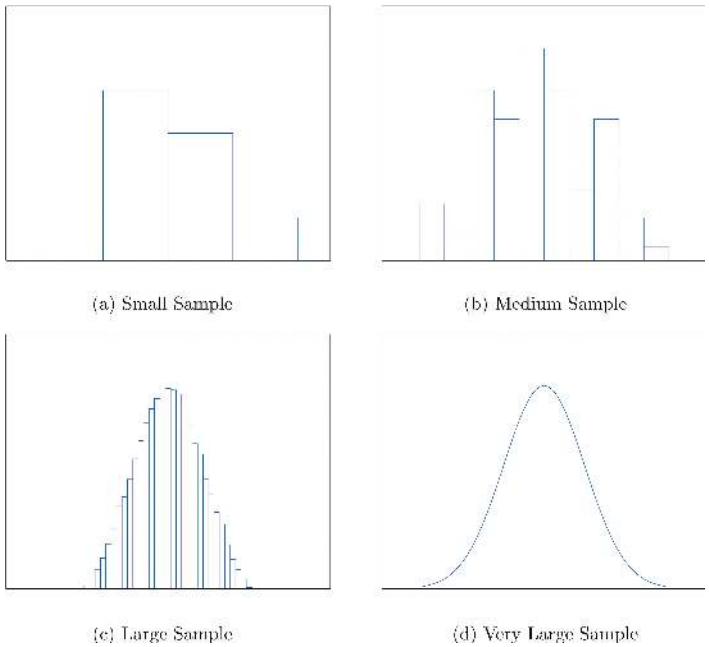


Figure 2.2.5: *Sample Size and Relative Frequency Histogram*

When the size n of a sample is small only a few classes can be used in constructing a relative frequency histogram. Such a histogram might look something like the one in panel (a) of Figure 2.2.5. If the sample size n were increased, then more classes

could be used in constructing a relative frequency histogram and the vertical bars of the resulting histogram would be finer, as indicated in panel (b) of Figure **2.2.5**. For a very large sample the relative frequency histogram would look very fine, like the one in (c) of Figure **2.2.5**. If the sample size were to increase indefinitely then the corresponding relative frequency histogram would be so fine that it would look like a smooth curve, such as the one in panel (d) of Figure **2.2.5**.

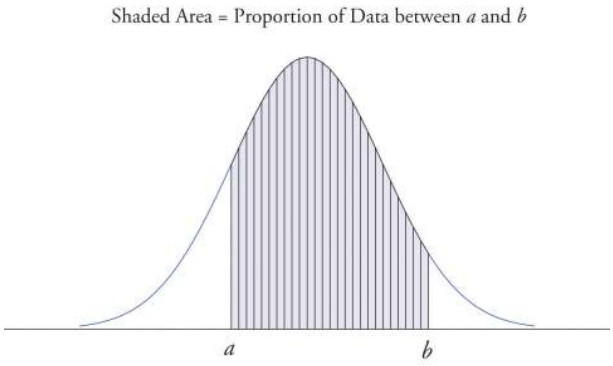


Figure **2.2.6**: *A Very Fine Relative Frequency Histogram*

It is common in statistics to represent a population or a very large data set by a smooth curve. It is good to keep in mind that such a curve is actually just a very fine relative frequency histogram in which the exceedingly narrow vertical bars have disappeared. Because the area of each such vertical bar is the proportion of the data that lies in the interval of numbers over which that bar stands, this means that for any two numbers a and b, the proportion of the data that lies between the two numbers a and b is the area under the curve that is above the interval (a, b) in the horizontal axis. This is the area shown in Figure **2.2.6**. In particular the total area under the curve is **1**, or **100%**.

Key Takeaway

- Graphical representations of large data sets provide a quick overview of the nature of the data.
- A population or a very large data set may be represented by a smooth curve. This curve is a very fine relative frequency histogram in which the exceedingly narrow vertical bars have been omitted.
- When a curve derived from a relative frequency histogram is used to describe a data set, the proportion of data with values between two numbers a and b is the area under the curve between a and b, as illustrated in Figure **2.2.6**.

This page titled 2.2: Histogram is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **2.1: Three Popular Data Displays** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

2.3: 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 2.3.1

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

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% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Exercise 2.3.2

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.

Answer

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

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 2.3.3

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?

HINT: 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.

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 2.3.4

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

Answer

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

The value 8.0 may be an outlier. Values appear to concentrate at one and two miles.

Example 2.3.5: Side-by-Side Stem-and-Leaf plot

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. Tables **2.3.1** and **2.3.2** show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Table **2.3.1**: Presidential Ages at Inauguration

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

2.3.2 *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		

Answer

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 1 1 1 1 1 0	5	3 6 6 7 7 8
9 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 3 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

? Exercise 2.3.6

The table shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Answer

Table **2.3.3**: Atlanta Hawks Wins and Losses

Number of Wins		Number of Losses
3	1	9
9 8 8 6 5	2	5 5 9
8 7 6 6 5 5 4 3 1 1 1 1 0	3	0 2 2 2 2 4 4 5 6 6 6 9 9 9
8 8 7 6 6 6 3 3 3 2 2 1 1 0	4	0 0 1 1 2 4 5 6 6 7 7 8 9
7 7 6 3 2 0 0 0 0	5	1 1 1 2 3 4 4 6 7
	6	9

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example, 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 2.3.7

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 and in Figure.

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

Answer

Figure **2.3.1**: A line graph showing the number of times a teenager needs to be reminded to do chores on the x-axis and frequency on the y-axis.

Exercise 2.3.8

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. Construct a line graph.

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

Answer

Figure **2.3.2**: A line graph showing the number of times a car is in the shop on the x-axis and frequency on the y-axis.

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 **2.3.9** has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 2.3.9

By the end of 2011, Facebook had over 146 million users in the United States. Table 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%

Answer

Figure **2.3.3**: This is a bar graph that matches the supplied data. The x-axis shows age groups and the y-axis show the percentages of Facebook users

Exercise 2.3.10

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

Answer

Figure **2.3.4**: This is a bar graph that matches the supplied data. The x-axis shows age groups, and the y-axis shows the percentages of Park City's population.

Example 2.3.11

The columns in Table 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%

Solution

Figure **2.3.5**: This is a bar graph that matches the supplied data. The x-axis shows race and ethnicity, and the y-axis shows the percentages of AP examinees.

Exercise 2.3.12

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.

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%

Answer

Figure **2.3.6**: This is a bar graph that matches the supplied data. The x-axis shows Park City voting districts, and the y-axis shows the percentages of the registered voter population.

Summary

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.

References

1. Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at www.kenburbary.com/2011/03/fa...-statistics-2/ (accessed August 21, 2013).
2. “9th Annual AP Report to the Nation.” CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goa...omoting-equity (accessed September 13, 2013).
3. “Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

This page titled 2.3: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

3: Descriptive Statistics

Statistics naturally divides into two branches, descriptive statistics and inferential statistics. Our main interest is in inferential statistics to try to infer from the data what the population might thin or to evaluate the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Nevertheless, the starting point for dealing with a collection of data is to organize, display, and summarize it effectively. These are the objectives of descriptive statistics, the topic of this chapter.

3.1: Measures of Center

3.2: Measures of Variability

3.3: Relative Position of Data

3.4: The Empirical Rule and Chebyshev's Theorem

This page titled 3: Descriptive Statistics is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

3.1: Measures of Center

Learning Objectives

- To learn the concept of the “center” of a data set.
- To learn the meaning of each of three measures of the center of a data set—the mean, the median, and the mode—and how to compute each one.

This section is be titled “three kinds of averages” because any kind of average could be used to answer the question "where is the center of the data?". We will see that the nature of the data set, as indicated by a relative frequency histogram, will determine what constitutes a good answer. Different shapes of the histogram call for different measures of central location.

The Mean

The first measure of central location is the usual “average” that is familiar to everyone: add up all the values, then divide by the number of values. Before writing a formula for the mean let us introduce some handy mathematical notation.

notations: $\sum$ "sum" and n "sample size"

The Greek letter $\sum$, pronounced "sigma", is a handy mathematical shorthand that stands for "add up all the values" or "sum". For example $\sum x$ means "add up all the values of x", and $\sum x^2$ means "add up all the values of x^2". In these expressions x usually stands for a value of the data, so $\sum x$ stands for "the sum of all the data values" and $\sum x^2$ means "the sum of the squares of all the data values".

n stands for the *sample size*, the number of data values. An example will help make this clear.

Example 3.1.1

Find n, $\sum x$, $\sum x^2$ and $\sum (x-1)^2$ for the data:

$$1,\ 3,\ 4$$

Solution

$$\begin{aligned} n &= 3 \quad \text{because there are three data values} \\ \sum x &= 1+3+4=8 \\ \sum x^2 &= 1^2+3^2+4^2=1+9+16=26 \\ \sum (x-1)^2 &= (1-1)^2+(3-1)^2+(4-1)^2=0^2+2^2+3^2=13 \end{aligned}$$

Using these handy notations it's easy to write a formula defining the mean $\bar{x}$ of a sample.

Definition: Sample Mean

The *sample mean* of a set of n sample data values is the number $\bar{x}$ defined by the formula

$$\bar{x} = \frac{\sum x}{n} \tag{3.1.1}$$

Example 3.1.2

Find the mean of the following sample data: $2,\ -1,\ 0,\ 2$

Solution

This is a application of Equation 3.1.1:

$$\bar{x} = \frac{\sum x}{n} = \frac{2+(-1)+0+2}{4} = \frac{3}{4} = 0.75$$

Example 3.1.3

A random sample of ten students is taken from the student body of a college and their GPAs are recorded as follows:

$$1.90, 3.00, 2.53, 3.71, 2.12, 1.76, 2.71, 1.39, 4.00, 3.33$$

Find the mean.

Solution

This is a application of Equation 3.1.1:

$$\bar{x} = \frac{\sum x}{n}$$
$$= \frac{1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.00 + 3.33}{10} = \frac{26.45}{10} = 2.645$$

Example 3.1.4

A random sample of 19 women beyond child-bearing age gave the following data, where x is the number of children and f is the frequency, or the number of times it occurred in the data set.

x	0	1	2	3	4
f	3	6	6	3	1

Find the sample mean.

Solution

In this example the data are presented by means of a data frequency table, introduced in Chapter 1. Each number in the first line of the table is a number that appears in the data set; the number below it is how many times it occurs. Thus the value 0 is observed three times, that is, three of the measurements in the data set are 0, the value 1 is observed six times, and so on. In the context of the problem this means that three women in the sample have had no children, six have had exactly one child, and so on. The explicit list of all the observations in this data set is therefore:

$$0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4$$

The sample size can be read directly from the table, without first listing the entire data set, as the sum of the frequencies: $n = 3 + 6 + 6 + 3 + 1 = 19$. The sample mean can be computed directly from the table as well:

$$\bar{x} = \frac{\sum x}{n} = \frac{0 \times 3 + 1 \times 6 + 2 \times 6 + 3 \times 3 + 4 \times 1}{19} = \frac{31}{19} = 1.6316$$

In the examples above the data sets were described as samples. Therefore the means were sample means $\bar{x}$. If the data come from a census, so that there is a measurement for every element of the population, then the mean is calculated by exactly the same process of summing all the measurements and dividing by how many of them there are, but it is now the ***population mean*** and is denoted by μ, the lower case Greek letter mu.

Definition: Population Mean

The *population mean* of a set of N population data is the number μ defined by the formula:

$$\mu = \frac{\sum x}{N}.$$

The mean of two numbers is the number that is halfway between them. For example, the average of the numbers 5 and 17 is $(5 + 17) / 2 = 11$, which is 6 units above 5 and 6 units below 17. In this sense the average 11 is the "center" of the data set $\{5, 17\}$. For larger data sets the mean can similarly be regarded as the "center" of the data.

The Median

To see why another concept of average is needed, consider the following situation. Suppose we are interested in the average yearly income of employees at a large corporation. We take a random sample of seven employees, obtaining the sample data (rounded to the nearest hundred dollars, and expressed in thousands of dollars).

$$24.8, 22.8, 24.6, 192.5, 25.2, 18.5, 23.7$$

The mean (rounded to one decimal place) is $\bar{x} = 47.4$, but the statement "the average income of employees at this corporation is $\$47,400$" is surely misleading. It is approximately twice what six of the seven employees in the sample make and is nowhere near what any of them makes. It is easy to see what went wrong: the presence of the one executive in the sample, whose salary is so large compared to everyone else's, caused the numerator in the formula for the sample mean to be far too large, pulling the mean far to the right of where we think that the average "ought" to be, namely around $\$24,000$ or $\$25,000$. The number 192.5 in our data set is called an outlier, a number that is far removed from most or all of the remaining measurements. Many times an outlier is the result of some sort of error, but not always, as is the case here. We would get a better measure of the "center" of the data if we were to arrange the data in numerical order:

$$18.5, 22.8, 23.7, 24.6, 24.8, 25.2, 192.5$$

then select the middle number in the list, in this case 24.6. The result is called the median of the data set, and has the property that roughly half of the measurements are larger than it is, and roughly half are smaller. In this sense it locates the center of the data. If there are an even number of measurements in the data set, then there will be two middle elements when all are lined up in order, so we take the mean of the middle two as the median. Thus we have the following definition.

Definition: Sample Median

The *sample median* $\tilde{x}$ of a set of sample data for which there are an odd number of measurements is the middle measurement when the data are arranged in numerical order.

The sample median of a set of sample data for which there are an even number of measurements, is the mean of the two middle measurements when the data are arranged in numerical order.

Definition: Population Median

The *population median* is defined in the same way as the sample median except for the entire population.

The median is a value that divides the observations in a data set so that 50% of the data are on its left and the other 50% on its right. In accordance with Figure $3.1.7$, therefore, in the curve that represents the distribution of the data, a vertical line drawn at the median divides the area in two, area 0.5 (50% of the total area 1) to the left and area 0.5 (50% of the total area 1) to the right, as shown in Figure $3.1.1$. In our income example the median, $\$24,600$, clearly gave a much better measure of the middle of the data set than did the mean $\$47,400$. This is typical for situations in which the distribution is skewed. (Skewness and symmetry of distributions are discussed at the end of this subsection.)

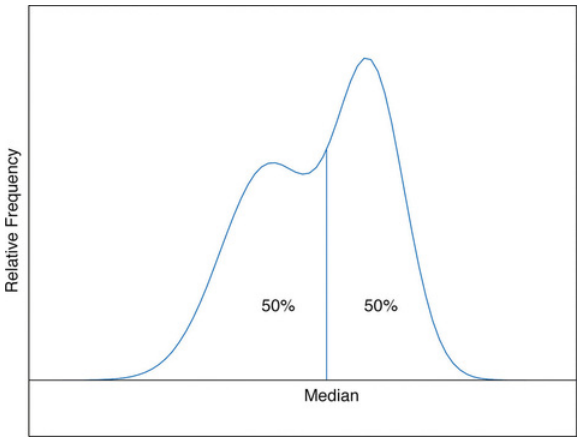


Figure 3.1.1: The Median

Example 3.1.5

Compute the sample median for the data from Example 3.1.2

Solution

The data in numerical order are $-1, 0, 2, 2$. The two middle measurements are 0 and 2, so $\tilde{x} = (0 + 2)/2 = 1$.

Example 3.1.6

Compute the sample median for the data from Example 3.1.3

Solution

The data in numerical order are

$$1.39, 1.76, 1.90, 2.12, 2.53, 2.71, 3.00, 3.33, 3.71, 4.00$$

The number of observations is ten, which is even, so there are two middle measurements, the fifth and sixth, which are 2.53 and 2.71. Therefore the median of these data is $\tilde{x} = (2.53 + 2.71)/2 = 2.62$.

Example 3.1.7

Compute the sample median for the data from Example 3.1.4

Solution

The data in numerical order are:

$$0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4$$

The number of observations is 19, which is odd, so there is one middle measurement, the tenth. Since the tenth measurement is 2, the median is $\tilde{x} = 2$.

In the last example it is important to note that we could have computed the median directly from the frequency table, without first explicitly listing all the observations in the data set. We already saw in Example 3.1.4 how to find the number of observations directly from the frequencies listed in the table $n = 3 + 6 + 6 + 3 + 1 = 19$. Thus the median is the tenth observation. The second line of the table in Example 3.1.4 shows that when the data are listed in order there will be three 0*s* followed by six 1*s*, so the tenth observation, the median, is 2.

The relationship between the mean and the median for several common shapes of distributions is shown in Figure 3.1.2. The distributions in panels (a) and (b) are said to be *symmetric* because of the symmetry that they exhibit. The distributions in the remaining two panels are said to be *skewed*. In each distribution we have drawn a vertical line that divides the area under the curve in half, which in accordance with Figure 3.1.1 is located at the median. The following facts are true in general:

- When the distribution is symmetric, as in panels (a) and (b) of Figure **3.1.2**, the mean and the median are equal.
- When the distribution is as shown in panel (c), it is said to be skewed right. The mean has been pulled to the right of the median by the long "right tail" of the distribution, the few relatively large data values.
- When the distribution is as shown in panel (d), it is said to be skewed left. The mean has been pulled to the left of the median by the long "left tail" of the distribution, the few relatively small data values.

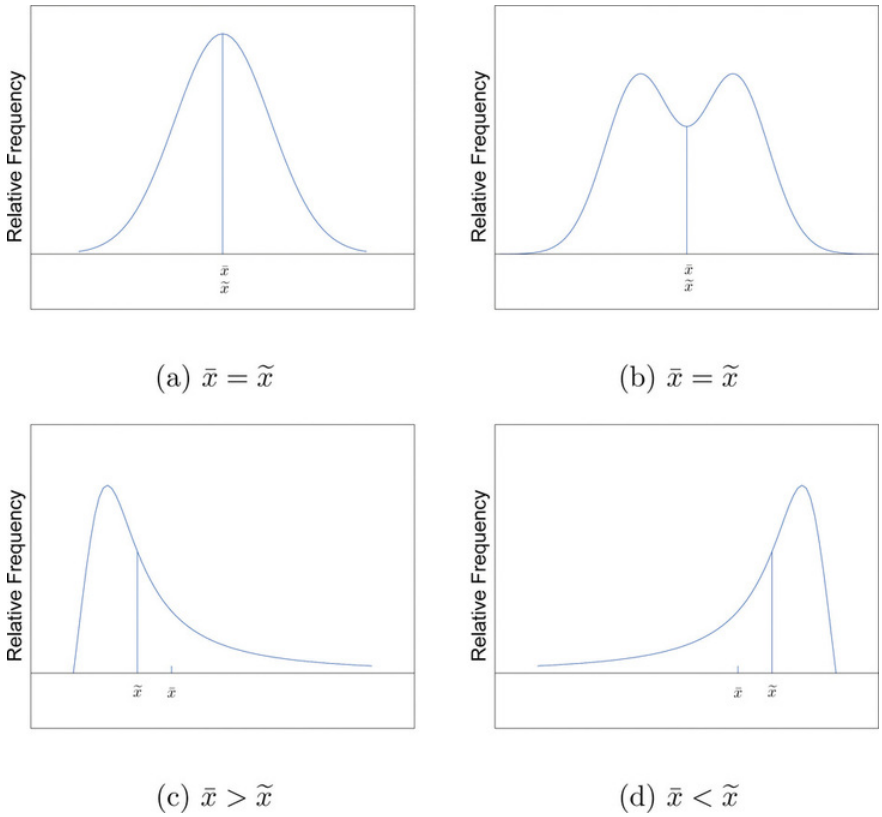


Figure **3.1.2**: Skewness of Relative Frequency Histograms

The Mode

Perhaps you have heard a statement like "The average number of automobiles owned by households in the United States is **1.37**," and have been amused at the thought of a fraction of an automobile sitting in a driveway. In such a context the following measure for central location might make more sense.

Definition: Sample Mode

The *sample mode* of a set of sample data is the most frequently occurring value.

On a relative frequency histogram, the highest point of the histogram corresponds to the mode of the data set. Figure **3.1.3** illustrates the mode.

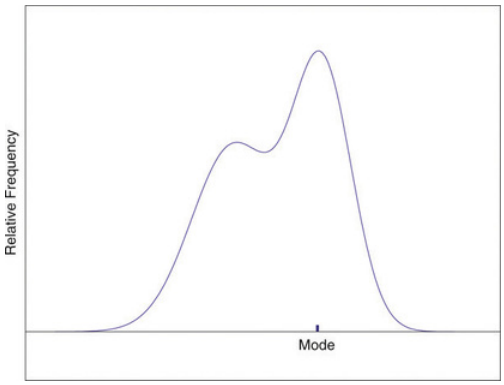


Figure* 3.1.3*: *Mode*

For any data set there is always exactly one mean and exactly one median. This need not be true of the mode; several different values could occur with the highest frequency, as we will see. It could even happen that every value occurs with the same frequency, in which case the concept of the mode does not make much sense.

Example 3.1.8

Find the mode of the following data set: $-1,\ 0,\ 2,\ 0$.

Solution

The value 0 is most frequently observed in the data set, so the mode is 0.

Example 3.1.9

Compute the sample mode for the data of Example 3.1.4

Solution

The two most frequently observed values in the data set are 1 and 2. Therefore mode is a set of two values: $\{1, 2\}$.

The mode is a measure of central location since most real-life data sets have more observations near the center of the data range and fewer observations on the lower and upper ends. The value with the highest frequency is often in the middle of the data range.

Key Takeaway

- The mean, the median, and the mode each answer the question “Where is the center of the data set?” The nature of the data set, as indicated by a relative frequency histogram, determines which one gives the best answer.

This page titled 3.1: Measures of Center is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **2.2: Measures of Central Location - Three Kinds of Averages** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

3.2: Measures of Variability

Learning Objectives

- To learn the concept of the variability of a data set.
- To learn how to compute three measures of the variability of a data set: the range, the variance, and the standard deviation.

Look at the two data sets in Table **3.2.1** and the graphical representation of each, called a *dot plot*, in Figure **3.2.1**.

Table **3.2.1**: *Two Data Sets*

Data Set I:	40	38	42	40	39	39	43	40	39	40
Data Set II:	46	37	40	33	42	36	40	47	34	45

The two sets of ten measurements each center at the same value: they both have mean, median, and mode **40**. Nevertheless a glance at the figure shows that they are markedly different. In Data Set I the measurements vary only slightly from the center, while for Data Set II the measurements vary greatly. Just as we have attached numbers to a data set to locate its center, we now wish to associate to each data set numbers that measure quantitatively how the data either scatter away from the center or cluster close to it. These new quantities are called measures of variability, and we will discuss three of them.

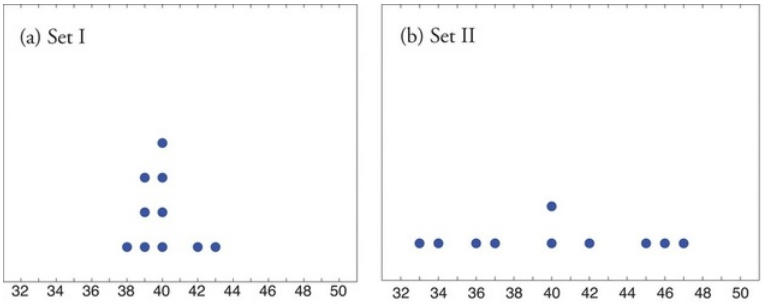


Figure **3.2.1**: Dot Plots of Data Sets

The Range

First we discuss the simplest measure of variability.

Definition: range

The *range* R of a data set is difference between its largest and smallest values

$$R = x_{\max} - x_{\min}$$

where $x_{\max}$ is the largest measurement in the data set and $x_{\min}$ is the smallest.

Example 3.2.1: Identifyig the Range of a dataset

Find the range of each data set in Table **3.2.1**.

Solution

- For Data Set I the maximum is **43** and the minimum is **38**, so the range is $R = 43 - 38 = 5$.
- For Data Set II the maximum is **47** and the minimum is **33**, so the range is $R = 47 - 33 = 14$.

The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite.

The Variance and the Standard Deviation

The other two measures of variability that we will consider are more elaborate and also depend on whether the data set is just a sample drawn from a much larger population or is the whole population itself (that is, a census).

Definition: sample variance and sample Standard Deviation

The *sample variance* of a set of n sample data is the number s^2 defined by the formula

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

which by algebra is equivalent to the formula

$$s^2 = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1}$$

The square root s of the sample variance is called the *sample standard deviation* of a set of n *sample data* . It is given by the formulas

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1}}.$$

Although the first formula in each case looks less complicated than the second, the latter is easier to use in hand computations, and is called a *shortcut formula*.

Example 3.2.2: Identifying the Variance and Standard Deviation of a Dataset

Find the sample variance and the sample standard deviation of Data Set II in Table 3.2.1.

Solution

To use the defining formula (the first formula) in the definition we first compute for each observation x its deviation $x - \bar{x}$ from the sample mean. Since the mean of the data is $\bar{x} = 40$, we obtain the ten numbers displayed in the second line of the supplied table

x	46	37	40	33	42	36	40	47	34	45
$x - \bar{x}$	−6	−3	0	−7	2	−4	0	7	−6	5

Thus

$$\sum (x - \bar{x})^2 = 6^2 + (-3)^2 + 0^2 + (-7)^2 + 2^2 + (-4)^2 + 0^2 + 7^2 + (-6)^2 + 5^2 = 224$$

so the variance is

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{224}{9} = 24.\bar{8}$$

and the standard deviation is

$$s = \sqrt{24.\bar{8}} \approx 4.99$$

The student is encouraged to compute the ten deviations for Data Set I and verify that their squares add up to 20, so that the sample variance and standard deviation of Data Set I are the much smaller numbers

$$s^2 = 20/9 = 2.\bar{2}$$

and

$$s = 20/9 \approx 1.49$$

Example 3.2.2

Find the sample variance and the sample standard deviation of the ten GPAs in "Example 2.2.3" in Section 2.2.

$$1.90 \;\; 3.00 \;\; 2.53 \;\; 3.71 \;\; 2.12 \;\; 1.76 \;\; 2.71 \;\; 1.39 \;\; 4.00 \;\; 3.33$$

Solution

Since

$$\sum x = 1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.00 + 3.33 = 26.45$$

and

$$\sum x^2 = 1.902 + 3.002 + 2.532 + 3.712 + 2.122 + 1.762 + 2.712 + 1.392 + 4.002 + 3.332 = 76.7321$$

the shortcut formula gives

$$s^2 = \frac{\sum x^2 - (\sum x)^2}{n-1} = \frac{76.7321 - (26.45)^2/10}{10-1} = \frac{6.77185}{9} = 0.752427\overline{7}$$

and

$$s = \sqrt{0.752427\overline{7}} \approx 0.867$$

The sample variance has different units from the data. For example, if the units in the data set were inches, the new units would be inches squared, or square inches. It is thus primarily of theoretical importance and will not be considered further in this text, except in passing.

If the data set comprises the whole population, then the *population* standard deviation, denoted σ (the lower case Greek letter sigma), and its square, the *population* variance σ^2, are defined as follows.

Definitions: The population variance σ^2 and population standard deviation σ

The variability of a set of N population data is measured by the population variance

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \tag{3.2.1}$$

and its square root, the population standard deviation

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \tag{3.2.2}$$

where μ is the population mean as defined above.

Note that the denominator in the fraction is the full number of observations, not that number reduced by one, as is the case with the sample standard deviation. Since most data sets are samples, we will always work with the sample standard deviation and variance.

Finally, in many real-life situations the most important statistical issues have to do with comparing the means and standard deviations of two data sets. Figure **3.2.2** illustrates how a difference in one or both of the sample mean and the sample standard deviation are reflected in the appearance of the data set as shown by the curves derived from the relative frequency histograms built using the data.

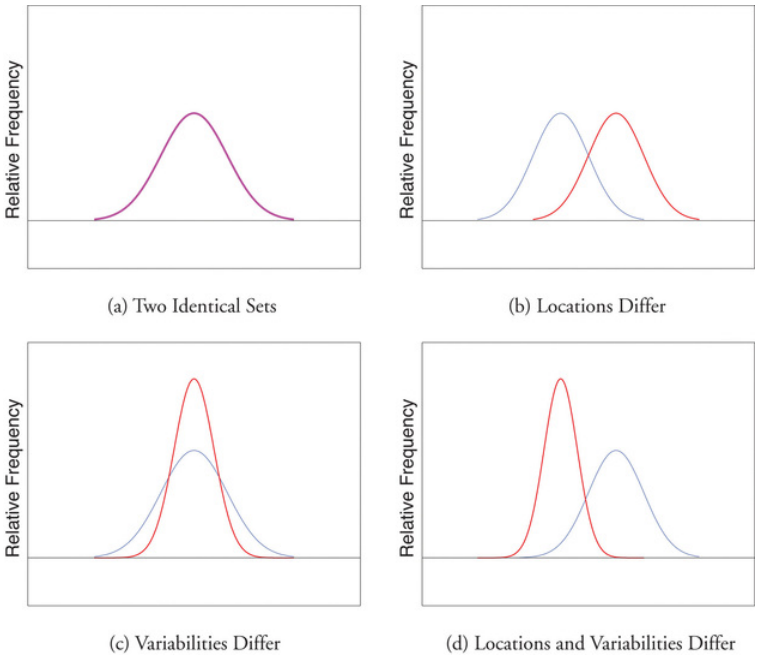


Figure **3.2.2**: Difference between Two Data Sets

Key Takeaway

The range, the standard deviation, and the variance each give a quantitative answer to the question "How variable are the data?"

This page titled 3.2: Measures of Variability is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **2.3: Measures of Variability** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

3.3: Relative Position of Data

Learning Objectives

- To learn the concept of the relative position of an element of a data set.
- To learn the meaning of each of two measures, the percentile rank and the z-score, of the relative position of a measurement and how to compute each one.
- To learn the meaning of the three quartiles associated to a data set and how to compute them.
- To learn the meaning of the five-number summary of a data set, how to construct the box plot associated to it, and how to interpret the box plot.

When you take an exam, what is often as important as your actual score on the exam is the way your score compares to other students' performance. If you made a 70 but the average score (whether the mean, median, or mode) was 85, you did relatively poorly. If you made a 70 but the average score was only 55 then you did relatively well. In general, the significance of one observed value in a data set strongly depends on how that value compares to the other observed values in a data set. Therefore we wish to attach to each observed value a number that measures its relative position.

Percentiles and Quartiles

Anyone who has taken a national standardized test is familiar with the idea of being given both a score on the exam and a "percentile ranking" of that score. You may be told that your score was 625 and that it is the 85th percentile. The first number tells how you actually did on the exam; the second says that 85% of the scores on the exam were less than or equal to your score, 625.

Definition: percentile of data

Given an observed value x in a data set, x is the Pth percentile of the data if $P\%$ of the data are less than or equal to x. The number P is the percentile rank of x.

Example 3.3.1

What percentile is the value 1.39 in the data set of ten GPAs considered in a previous Example? What percentile is the value 3.33?

Solution

The data, written in increasing order, are

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 4.00$$

The only data value that is less than or equal to 1.39 is 1.39 itself. Since 1 out of ten, or $1/10 = 10\%$, of the data points are less than or equal to 1.39, 1.39 is the 10th percentile. Eight data values are less than or equal to 3.33. Since 8 out of ten, or $8 / 10 = .80 = 80\%$ of the data values are less than or equal to 3.33, the value 3.33 is the 80th percentile of the data.

The Pth percentile cuts the data set in two so that approximately $P\%$ of the data lie below it and $(100 - P)\%$ of the data lie above it. In particular, the three percentiles that cut the data into fourths, as shown in Figure $3.3.1$, are called the quartiles of a data set. The quartiles are the three numbers Q_1, Q_2, Q_3 that divide the data set approximately into fourths. The following simple computational definition of the three quartiles works well in practice.

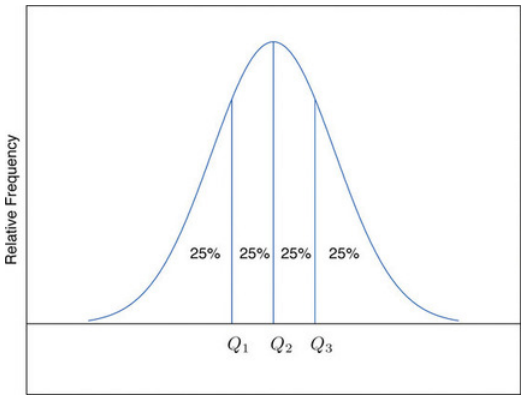


Figure 3.3.1: Data Division by Quartiles

Definition: quartile

For any data set:

1. The second quartile Q_2 of the data set is its median.
2. Define two subsets:
    1. the lower set: all observations that are strictly less than Q_2
    2. the upper set: all observations that are strictly greater than Q_2
3. The first quartile Q_1 of the data set is the median of the lower set.
4. The third quartile Q_3 of the data set is the median of the upper set.

Example 3.3.2

Find the quartiles of the data set of GPAs of discussed in a previous Example.

Solution

As in the previous example we first list the data in numerical order:

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 4.00$$

This data set has $n = 10$ observations. Since 10 is an even number, the median is the mean of the two middle observations:

$$\tilde{x} = (2.53 + 2.71) \,/\, 2 = 2.62.$$

Thus the second quartile is $Q_2 = 2.62$. The lower and upper subsets are

- Lower: $L = \{1.39, 1.76, 1.90, 2.12, 2.53\}$,
- Upper: $U = \{2.71, 3.00, 3.33, 3.71, 4.00\}$.

Each has an odd number of elements, so the median of each is its middle observation. Thus the first quartile is $Q_1 = 1.90$, the median of L, and the third quartile is $Q_3 = 3.33$, the median of U.

Example 3.3.3

Adjoin the observation 3.88 to the data set of the previous example and find the quartiles of the new set of data.

Solution

As in the previous example we first list the data in numerical order:

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 3.88, 4.00$$

This data set has 11 observations. The second quartile is its median, the middle value 2.71. Thus $Q_2 = 2.71$. The lower and upper subsets are now

- Lower: $L = \{1.39, 1.76, 1.90, 2.12, 2.53\}$

- Upper: $U = \{3.00, 3.33, 3.71, 3.88, 4.00\}$.

The lower set L has median, the middle value, 1.90, so $Q_1 = 1.90$. The upper set has median 3.71, so $Q_3 = 3.71$.

In addition to the three quartiles, the two extreme values, the minimum x_{min} and the maximum x_{max} are also useful in describing the entire data set. Together these five numbers are called the five-number summary of a data set,

$$\{ X_{min}, Q_1, Q_2, Q_3, X_{max} \}$$

The five-number summary is used to construct a box plot, as in Figure 3.3.2. Each of the five numbers is represented by a vertical line segment, a box is formed using the line segments at Q_1 and Q_3 as its two vertical sides, and two horizontal line segments are extended from the vertical segments marking Q_1 and Q_3 to the adjacent extreme values. (The two horizontal line segments are referred to as “whiskers,” and the diagram is sometimes called a "box and whisker plot.") We caution the reader that there are other types of box plots that differ somewhat from the ones we are constructing, although all are based on the three quartiles.

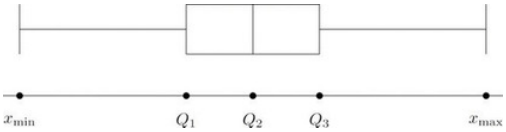


Figure 3.3.2: The Box Plot

Note that the distance from $Q1$ to $Q3$ is the length of the interval over which the middle half of the data range. Thus it has the following special name.

Definition: interquartile range

The interquartile range IQR is the quantity

$$IQR = Q3 - Q1$$

Example 3.3.4

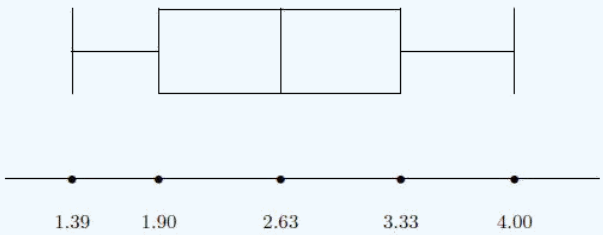
Construct a box plot and find the IQR for the data in Example 3.3.3.

Solution

From our work in Example 3.3.1, we know that the five-number summary is

- $X_{min} = 1.39$
- $Q1 = 1.90$
- $Q2 = 2.62$
- $Q3 = 3.33$
- $X_{max} = 4.00$

The box plot is:



The interquartile range is: $IQR = 3.33 - 1.90 = 1.43$.

z-Scores

Another way to locate a particular observation x in a data set is to compute its distance from the mean in units of standard deviation. The z-score indicates how many standard deviations an individual observation x is from the center of the data set, its

mean. It is used on distributions that have been *standardized*, which allows us to better understand its properties. If z is negative then x is below average. If z is 0 then x is equal to the average. If z is positive then x is above the average

Definition: z-score

The z-score of an observation x is the number z given by the computational formula

$$z = \frac{x - \mu}{\sigma}$$

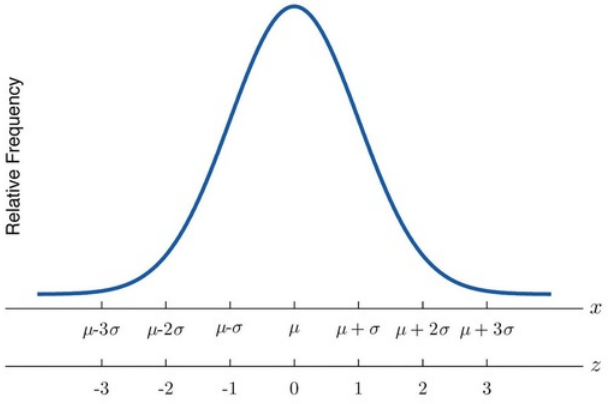


***Figure* 3.3.3:** *x*-Scale versus *z*-Score

Example 3.3.5

Suppose the mean and standard deviation of the GPA's of all currently registered students at a college are $\mu = 2.70$ and $\sigma = 0.50$. The z-scores of the GPA's of two students, Antonio and Beatrice, are $z = -0.62$ and $z = 1.28$, respectively. What are their GPAs?

Solution

Using the second formula right after the definition of z-scores we compute the GPA's as

- Antonio: $x = \mu + z\sigma = 2.70 + (-0.62)(0.50) = 2.39$
- Beatrice: $x = \mu + z\sigma = 2.70 + (1.28)(0.50) = 3.34$

Key Takeaways

- The percentile rank and z-score of a measurement indicate its relative position with regard to the other measurements in a data set.
- The three quartiles divide a data set into fourths.
- The five-number summary and its associated box plot summarize the location and distribution of the data.

This page titled 3.3: Relative Position of Data is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **2.4: Relative Position of Data** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

3.4: The Empirical Rule and Chebyshev's Theorem

Learning Objectives

- To learn what the value of the standard deviation of a data set implies about how the data scatter away from the mean as described by the *Empirical Rule* and *Chebyshev's Theorem*.
- To use the Empirical Rule and Chebyshev's Theorem to draw conclusions about a data set.

You probably have a good intuitive grasp of what the average of a data set says about that data set. In this section we begin to learn what the standard deviation has to tell us about the nature of the data set.

The Empirical Rule

We start by examining a specific set of data. Table **3.4.1** shows the heights in inches of **100** randomly selected adult men. A relative frequency histogram for the data is shown in Figure **3.4.1**. The mean and standard deviation of the data are, rounded to two decimal places, $\bar{x} = 69.92$ and $\sigma = 1.70$.

Table **3.4.1**: Heights of Men

68.7	72.3	71.3	72.5	70.6	68.2	70.1	68.4	68.6	70.6
73.7	70.5	71.0	70.9	69.3	69.4	69.7	69.1	71.5	68.6
70.9	70.0	70.4	68.9	69.4	69.4	69.2	70.7	70.5	69.9
69.8	69.8	68.6	69.5	71.6	66.2	72.4	70.7	67.7	69.1
68.8	69.3	68.9	74.8	68.0	71.2	68.3	70.2	71.9	70.4
71.9	72.2	70.0	68.7	67.9	71.1	69.0	70.8	67.3	71.8
70.3	68.8	67.2	73.0	70.4	67.8	70.0	69.5	70.1	72.0
72.2	67.6	67.0	70.3	71.2	65.6	68.1	70.8	71.4	70.2
70.1	67.5	71.3	71.5	71.0	69.1	69.5	71.1	66.8	71.8
69.6	72.7	72.8	69.6	65.9	68.0	69.7	68.7	69.8	69.7

If we go through the data and count the number of observations that are within one standard deviation of the mean, that is, that are between $69.92 - 1.70 = 68.22$ and $69.92 + 1.70 = 71.62$ inches, there are **69** of them. If we count the number of observations that are within two standard deviations of the mean, that is, that are between $69.92 - 2(1.70) = 66.52$ and $69.92 + 2(1.70) = 73.32$ inches, there are **95** of them. All of the measurements are within three standard deviations of the mean, that is, between $69.92 - 3(1.70) = 64.822$ and $69.92 + 3(1.70) = 75.02$ inches. These tallies are not coincidences, but are in agreement with the following result that has been found to be widely applicable.

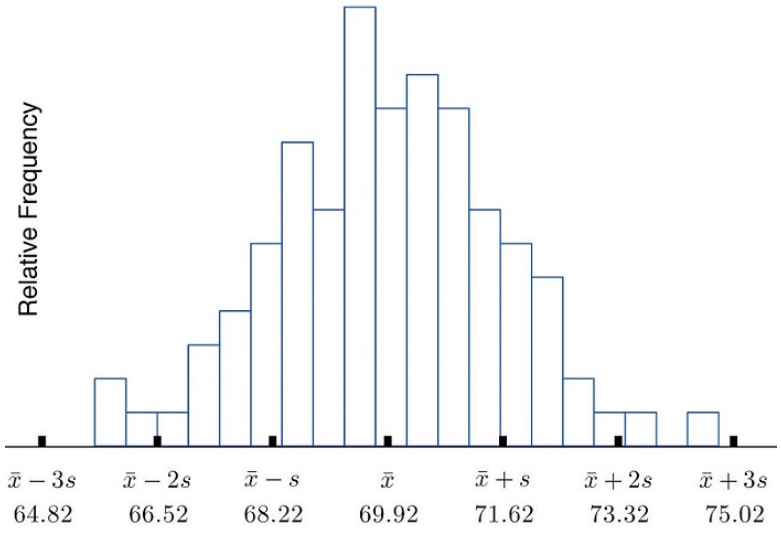


Figure 3.4.1: Heights of Adult Men

The Empirical Rule

Approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\bar{x} \pm s$ for samples and with endpoints $\mu \pm \sigma$ for populations; if a data set has an approximately bell-shaped relative frequency histogram, then (Figure 3.4.2)

- approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations; and
- approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.

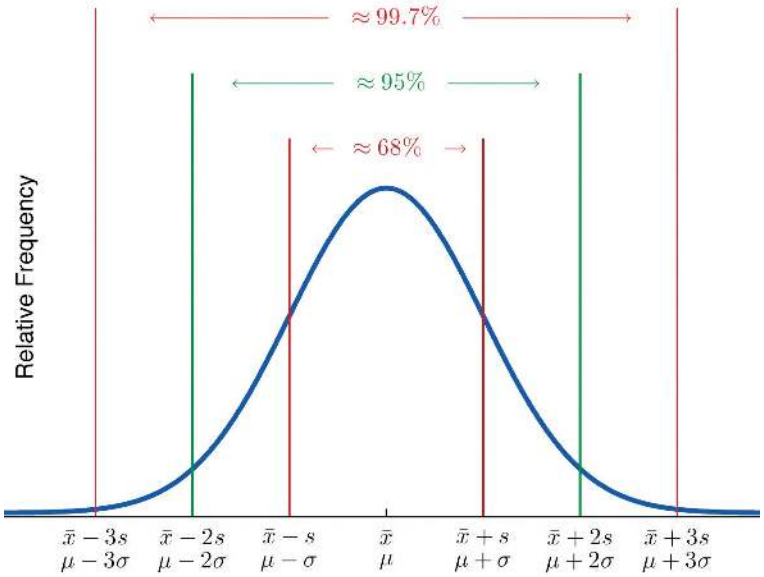


Figure 3.4.2: The Empirical Rule

Two key points in regard to the Empirical Rule are that the data distribution must be approximately bell-shaped and that the percentages are only approximately true. The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule. We see this with the example of the heights of the men: the Empirical Rule suggested 68 observations between 68.22 and 71.62 inches, but we counted 69.

Example 3.4.1

Heights of 18-year-old males have a bell-shaped distribution with mean 69.6 inches and standard deviation 1.4 inches.

1. About what proportion of all such men are between 68.2 and 71 inches tall?
2. What interval centered on the mean should contain about 95% of all such men?

Solution

A sketch of the distribution of heights is given in Figure 3.4.3.

1. Since the interval from 68.2 to 71.0 has endpoints $\bar{x} - s$ and $\bar{x} + s$, by the Empirical Rule about 68% of all 18-year-old males should have heights in this range.
2. By the Empirical Rule the shortest such interval has endpoints $\bar{x} - 2s$ and $\bar{x} + 2s$. Since

$$\bar{x} - 2s = 69.6 - 2(1.4) = 66.8$$

and

$$\bar{x} + 2s = 69.6 + 2(1.4) = 72.4$$

the interval in question is the interval from 66.8 inches to 72.4 inches.

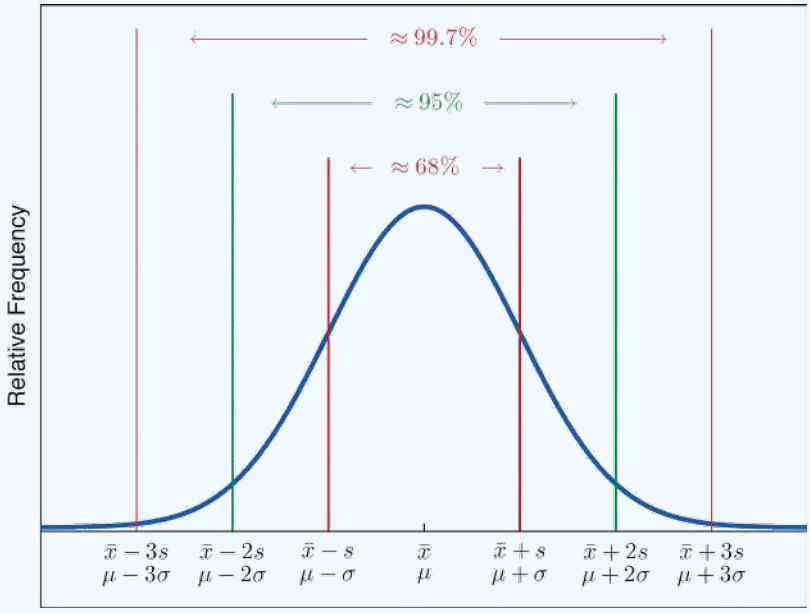


Figure 3.4.3: Distribution of Heights

Example 3.4.2

Scores on IQ tests have a bell-shaped distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. Discuss what the Empirical Rule implies concerning individuals with IQ scores of 110, 120, and 130.

Solution

A sketch of the IQ distribution is given in Figure 3.4.3. The Empirical Rule states that

1. approximately 68% of the IQ scores in the population lie between 90 and 110,
2. approximately 95% of the IQ scores in the population lie between 80 and 120, and
3. approximately 99.7% of the IQ scores in the population lie between 70 and 130.

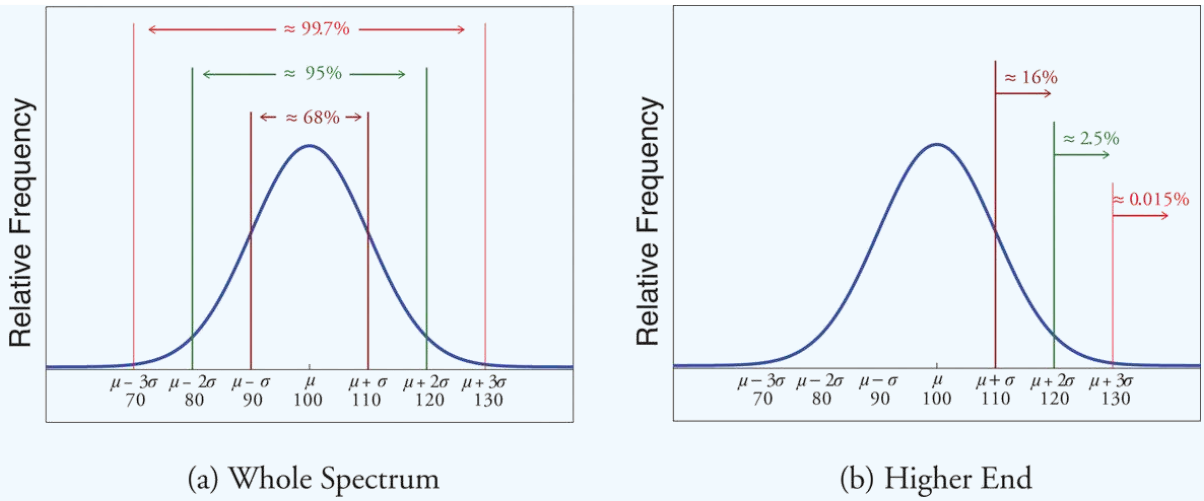


Figure 3.4.3: Distribution of IQ Scores.

1. Since 68% of the IQ scores lie *within* the interval from 90 to 110, it must be the case that 32% lie *outside* that interval. By symmetry approximately half of that 32%, or 16% of all IQ scores, will lie above 110. If 16% lie above 110, then 84% lie below. We conclude that the IQ score 110 is the 84th percentile.
2. The same analysis applies to the score 120. Since approximately 95% of all IQ scores lie within the interval form 80 to 120, only 5% lie outside it, and half of them, or 2.5% of all scores, are above 120. The IQ score 120 is thus higher than 97.5% of all IQ scores, and is quite a high score.
3. By a similar argument, only 15/100 of 1% of all adults, or about one or two in every thousand, would have an IQ score above 130. This fact makes the score 130 extremely high.

Chebyshev's Theorem

The Empirical Rule does not apply to all data sets, only to those that are bell-shaped, and even then is stated in terms of approximations. A result that applies to every data set is known as Chebyshev's Theorem.

Chebyshev's Theorem

For any numerical data set,

- at least 3/4 of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations;
- at least 8/9 of the data lie within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations;
- at least $1 - 1/k^2$ of the data lie within k standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm ks$ for samples and with endpoints $\mu \pm k\sigma$ for populations, where k is any positive whole number that is greater than 1.

Figure 3.4.4 gives a visual illustration of Chebyshev's Theorem.

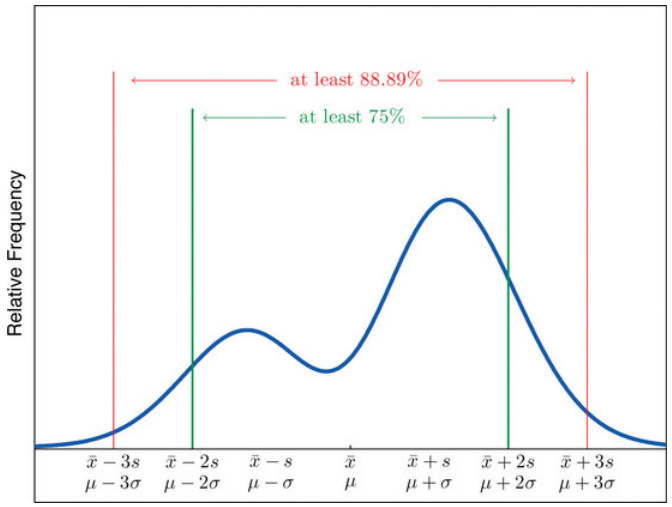


Figure **3.4.4**: Chebyshev's Theorem

It is important to pay careful attention to the words **"at least"** at the beginning of each of the three parts of Chebyshev's Theorem. The theorem gives the *minimum* proportion of the data which must lie within a given number of standard deviations of the mean; the true proportions found within the indicated regions could be greater than what the theorem guarantees.

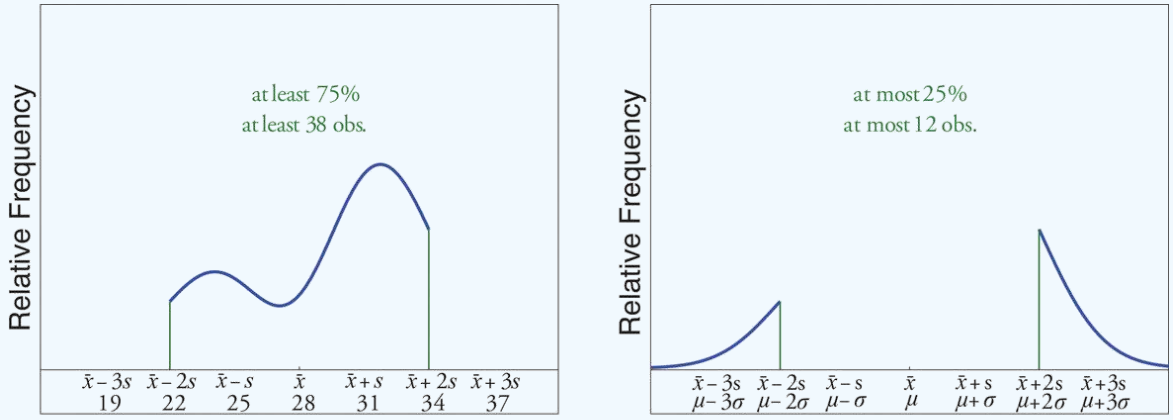
Example 3.4.3

A sample of size $n = 50$ has mean $\bar{x} = 28$ and standard deviation $s = 3$. Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval $(22, 34)$? What can be said about the number of observations that lie outside that interval?

Solution

The interval $(22, 34)$ is the one that is formed by adding and subtracting two standard deviations from the mean. By Chebyshev's Theorem, at least $3/4$ of the data are within this interval. Since $3/4$ of 50 is 37.5, this means that at least 37.5 observations are in the interval. But one cannot take a fractional observation, so we conclude that at least 38 observations must lie inside the interval $(22, 34)$.

If at least $3/4$ of the observations are in the interval, then at most $1/4$ of them are outside it. Since $1/4$ of 50 is 12.5, at most 12.5 observations are outside the interval. Since again a fraction of an observation is impossible, x $(22, 34)$.



(a) Within $\bar{x} \pm 2s$ (b) Outside $\bar{x} \pm 2$

Example 3.4.4

The number of vehicles passing through a busy intersection between $8:00\ a.m.$ and $10:00\ a.m.$ was observed and recorded on every weekday morning of the last year. The data set contains $n = 251$ numbers. The sample mean is $\bar{x} = 725$ and the sample standard deviation is $s = 25$. Identify which of the following statements *must* be true.

1. On approximately 95% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
2. On at least 75% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
3. On at least 189 weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
4. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was either less than 675 or greater than 775.
5. On at most 12.5% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.
6. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.

Solution

1. Since it is not stated that the relative frequency histogram of the data is bell-shaped, the Empirical Rule does not apply. Statement (1) is based on the Empirical Rule and therefore it might not be correct.
2. Statement (2) is a direct application of part (1) of Chebyshev's Theorem because $\bar{x} - 2s$, $\bar{x} + 2s = (675, 775)$. It must be correct.
3. Statement (3) says the same thing as statement (2) because 75% of 251 is 188.25, so the minimum whole number of observations in this interval is 189. Thus statement (3) is definitely correct.
4. Statement (4) says the same thing as statement (2) but in different words, and therefore is definitely correct.
5. Statement (4), which is definitely correct, states that at most 25% of the time either fewer than 675 or more than 775 vehicles passed through the intersection. Statement (5) says that half of that 25% corresponds to days of light traffic. This would be correct if the relative frequency histogram of the data were known to be symmetric. But this is not stated; perhaps all of the observations outside the interval $(675, 775)$ are less than 75. Thus statement (5) might not be correct.
6. Statement (4) is definitely correct and statement (4) implies statement (6): even if every measurement that is outside the interval $(675, 775)$ is less than 675 (which is conceivable, since symmetry is not known to hold), even so at most 25% of all observations are less than 675. Thus statement (6) must definitely be correct.

Key Takeaway

- The Empirical Rule is an approximation that applies only to data sets with a bell-shaped relative frequency histogram. It estimates the proportion of the measurements that lie within one, two, and three standard deviations of the mean.
- Chebyshev's Theorem is a fact that applies to all possible data sets. It describes the minimum proportion of the measurements that lie must within one, two, or more standard deviations of the mean.

This page titled 3.4: The Empirical Rule and Chebyshev's Theorem is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **2.5: The Empirical Rule and Chebyshev's Theorem** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

CHAPTER OVERVIEW

4: Probability Topics

Probability theory is concerned with probability, the analysis of random phenomena. The central objects of probability theory are random variables, stochastic processes, and events: mathematical abstractions of non-deterministic events or measured quantities that may either be single occurrences or evolve over time in an apparently random fashion.

- 4.1: Introduction
  - 4.1.1: Terminology
  - 4.1.2: Independent and Mutually Exclusive Events
- 4.2: Addition and Multiplication Rule of Probability
- 4.3: Conditional Probability using Contingency Tables
- 4.E: Probability Topics (Exericses)

This page titled 4: Probability Topics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.1: Introduction



Figure 4.1.1. Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

CHAPTER OBJECTIVES

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

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

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.

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.

Collaborative Exercise

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $P(\text{change})$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $P(\text{bus})$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $P(\text{change})$.
- Find $P(\text{bus})$.
- Find $P(\text{change AND bus})$. Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $P(\text{change}|\text{bus})$. Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

This page titled 4.1: Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.1.1: 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. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: 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 $P(A)$.

Definition: Probability

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. 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 = rolling a number that is at least five. There are two outcomes {5, 6}. $P(E) = \frac{2}{6}$. 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.

Definition: Law of Large Numbers

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.

The "OR" Event

An outcome is in the event **A OR B** if the outcome is in **A** or is in **B** or is in both **A** and **B**. For example, let $\mathbf{A} = \{1, 2, 3, 4, 5\}$ and $\mathbf{B} = \{4, 5, 6, 7, 8\}$. **A OR B** $= \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

The "AND" Event

An outcome is in the event **A AND 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 AND B** $= 4, 5$.

The **complement** of event **A** is denoted **A'** (read "A prime"). **A'** consists of all outcomes that are **NOT** in **A**. Notice that

$$P(\mathrm{A}) + P(\mathrm{A}') = 1.$$

For example, let $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$ and let $\mathbf{A} = 1, 2, 3, 4$. Then, $\mathbf{A}' = 5, 6$ and $P(A) = \frac{4}{6}$, $P(\mathrm{A}') = \frac{2}{6}$, and

$$P(\mathrm{A}) + P(\mathrm{A}') = \frac{4}{6} + \frac{2}{6} = 1.$$

The conditional probability of **A** given **B** is written $P(\mathrm{A}|\mathrm{B})$. $P(\mathrm{A}|\mathrm{B})$ 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 $P(\mathrm{A}|\mathrm{B})$ is

$$P(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}}$$

where $P(\mathrm{B})$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$. Let **A** = face is 2 or 3 and **B** = face is even (2, 4, 6). To calculate $P(\mathrm{A}|\mathrm{B})$, we count the number of outcomes 2 or 3 in the sample space $\mathbf{B} = \{2, 4, 6\}$. 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.

$$\begin{aligned} P(\mathrm{A}|\mathrm{B}) &= \frac{\mathrm{P(A\ AND\ B)}}{P(\mathrm{B})} \\ &= \frac{\frac{\text{the number of outcomes that are 2 or 3 and even in S}}{6}}{\frac{\text{the number of outcomes that are even in S}}{6}} \\ &= \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3} \end{aligned}$$

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 4.1.1.1

The sample space S is the whole numbers starting at one and less than 20.

a. $S =$ _____

Let event $A =$ the even numbers and event $B =$ numbers greater than 13.

b. $A =$ _____, $B =$ _____
c. $P(\text{A}) =$ _____, $P(\text{B}) =$ _____
d. A AND B = _____, A OR B = _____
e. $P(\text{A AND B}) =$ _____, $P(\text{A OR B}) =$ _____
f. $\text{A}' =$ _____, $P(\text{A}') =$ _____
g. $P(\text{A}) + P(\text{A}') =$ _____
h. $P(\text{A|B}) =$ _____, $P(\text{B|A}) =$ _____; are the probabilities equal?

Answer

a. $\text{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
b. $\text{A} = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, \text{B} = \{14, 15, 16, 17, 18, 19\}$
c. $P(\text{A}) = \frac{9}{19}$, $P(\text{B}) = \frac{6}{19}$
d. A AND B $= \{14, 16, 18\}$, A OR B $= \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
e. $P(\text{A AND B}) = \frac{3}{19}$, $P(\text{A OR B}) = \frac{12}{19}$
f. $\text{A}' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(\text{A}') = \frac{10}{19}$
g. $P(\text{A}) + P(\text{A}') = 1 \left(\left(\frac{9}{19} + \frac{10}{19} = 1\right)\right.$
h. $P(\text{A|B}) = \frac{\text{P(A AND B)}}{\text{P(B)}} = \frac{3}{6}, P(\text{B|A}) = \frac{\text{P(A AND B)}}{\text{P(A)}} = \frac{3}{9}$, No

Exercise 4.1.1.1

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. $S =$ _____

Let event $A =$ the sum is even and event $B =$ the first number is prime.

b. $A =$ _____, $B =$ _____
c. $P(\text{A}) =$ _____, $P(\text{B}) =$ _____
d. A AND B = _____, A OR B = _____
e. $P(\text{A AND B}) =$ _____, $P(\text{A OR B}) =$ _____
f. $\text{B}' =$ _____, $P(\text{B}') =$ _____
g. $P(\text{A}) + P(\text{A}') =$ _____
h. $P(\text{A|B}) =$ _____, $P(\text{B|A}) =$ _____; are the probabilities equal?

Answer

a. $\text{S} = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
b. $\text{A} = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)\}$
$\text{B} = \{(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
c. $P(\text{A}) = \frac{1}{2}$, $P(\text{B}) = \frac{2}{3}$
d. A AND B $= \{(2,2), (2,4), (3,1), (3,3)\}$
A OR B $= \{(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
e. $P(\text{A AND B}) = \frac{1}{3}, P(\text{A OR B}) = \frac{5}{6}$
f. $\text{B}' = \{(1,1), (1,2), (1,3), (1,4)\}, P(\text{B}') = \frac{1}{3}$
g. $P(\text{B}) + P(\text{B}') = 1$
h. $P(\text{A|B}) = \frac{P(\text{A AND B})}{P(\text{B})} = \frac{1}{2}, P(\text{B|A}) = \frac{P(\text{A AND B})}{P(\text{B})} = \frac{2}{3}$, No.

Example 4.1.1.2A

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).

a. Event $\mathbf{T}$ = the outcome is two.
b. Event $\mathbf{A}$ = the outcome is an even number.
c. Event $\mathbf{B}$ = the outcome is less than four.
d. The complement of $\mathbf{A}$.
e. $\mathbf{A}$ GIVEN $\mathbf{B}$
f. $\mathbf{B}$ GIVEN $\mathbf{A}$
g. $\mathbf{A}$ AND $\mathbf{B}$
h. $\mathbf{A}$ OR $\mathbf{B}$
i. $\mathbf{A}$ OR $\mathbf{B}'$
j. Event $\mathbf{N}$ = the outcome is a prime number.
k. Event $\mathbf{I}$ = the outcome is seven.

Solution

a. $\mathbf{T} = \{2\}$, $P(\mathbf{T}) = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(\mathbf{A}) = \frac{1}{2}$
c. $\mathbf{B} = \{1, 2, 3\}$, $P(\mathbf{B}) = \frac{1}{2}$
d. $\mathbf{A}' = \{1, 3, 5\}$, $P(\mathbf{A}') = \frac{1}{2}$
e. $\mathbf{A}|\mathbf{B} = \{2\}$, $P(\mathbf{A}|\mathbf{B}) = \frac{1}{3}$
f. $\mathbf{B}|\mathbf{A} = \{2\}$, $P(\mathbf{B}|\mathbf{A}) = \frac{1}{3}$
g. $\mathbf{A}$ AND $\mathbf{B} = 2$, $P(\mathbf{A}$ AND $\mathbf{B}) = \frac{1}{6}$
h. $\mathbf{A}$ OR $\mathbf{B} = \{1, 2, 3, 4, 6\}$, $P(\mathbf{A}$ OR $\mathbf{B}) = \frac{5}{6}$
i. $\mathbf{A}$ OR $\mathbf{B}' = \{2, 4, 5, 6\}$, $P(\mathbf{A}$ OR $\mathbf{B}') = \frac{2}{3}$
j. $\mathbf{N} = \{2, 3, 5\}$, $P(\mathbf{N}) = \frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7) = 0$.

Example 4.1.1.2B

Table describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
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:

a. $P(\mathbf{M})$
b. $P(\mathbf{F})$
c. $P(\mathbf{R})$
d. $P(\mathbf{L})$
e. $P(\mathbf{M}$ AND $\mathbf{R})$
f. $P(\mathbf{F}$ AND $\mathbf{L})$
g. $P(\mathbf{M}$ OR $\mathbf{F})$
h. $P(\mathbf{M}$ OR $\mathbf{R})$
i. $P(\mathbf{F}$ OR $\mathbf{L})$

j. $P(\text{M'})$
k. $P(\text{R}|\text{M})$
l. $P(\text{F}|\text{L})$
m. $P(\text{L}|\text{F})$

Answer

a. $P(\text{M}) = 0.52$
b. $P(\text{F}) = 0.48$
c. $P(\text{R}) = 0.87$
d. $P(\text{L}) = 0.13$
e. $P(\text{M AND R}) = 0.43$
f. $P(\text{F AND L}) = 0.04$
g. $P(\text{M OR F}) = 1$
h. $P(\text{M OR R}) = 0.96$
i. $P(\text{F OR L}) = 0.57$
j. $P(\text{M'}) = 0.48$
k. $P(\text{R}|\text{M}) = 0.8269$ (rounded to four decimal places)
l. $P(\text{F}|\text{L}) = 0.3077$ (rounded to four decimal places)
m. $P(\text{L}|\text{F}) = 0.0833$

References

1. "Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Review

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.

Formula Review

A and **B** are events

$P(\text{S}) = 1$ where **S** is the sample space

$0 \leq P(\text{A}) \leq 1$

$P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{\text{P(B)}}$

Glossary

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

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 letters such as A, B, C, and so on.

Experiment

a planned activity carried out under controlled conditions

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:

- $0 \leq P(\text{A}) \leq 1$
- If A and B are any two mutually exclusive events, then $\text{P(A OR B)} = P(\text{A}) + P(\text{B})$.
- $P(\text{S}) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of *A* GIVEN *B*

$P(\text{A}|\text{B})$ is the probability that event A will occur given that the event B has already occurred.

The Or Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

? Exercise 3.2.2

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.

Answer

a. $P(\text{L}') = P(\text{S})$
b. $P(\text{M OR S})$
c. $P(\text{F AND L})$
d. $P(\text{M}|\text{L})$

e. $P(\text{L}|\text{M})$
f. $P(\text{S}|\text{F})$
g. $P(\text{F}|\text{L})$
h. $P(\text{F OR L})$
i. $P(\text{M AND S})$
j. $P(\text{F})$

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.

Exercise 3.2.3

Find $P(\text{H})$.

Exercise 3.2.4

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise 3.2.5

Find $P(\text{F})$.

Exercise 3.2.6

Find $P(\text{C})$.

Answer

$P(\text{C}) = \frac{5}{42} = 0.12$

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.

Exercise 3.2.7

Find $P(\mathrm{B})$.

Exercise 3.2.8

Find $P(\mathrm{G})$.

Answer

$P(\mathrm{G}) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise 3.2.9

Find $P(\mathrm{P})$.

Exercise 3.2.10

Find $P(\mathrm{R})$.

Answer

$P(\mathrm{R}) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise 3.2.11

Find $P(\mathrm{Y})$.

Exercise 3.2.12

Find $P(\mathrm{O})$.

Answer

$P(textO) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

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.

Exercise 3.2.13

Find $P(\mathrm{A})$.

Exercise 3.2.14

Find $P(\mathrm{E})$.

Answer

$P(\text{E}) = \frac{47}{194} = 0.24$

Exercise 3.2.15

Find $P(\text{F})$.

Exercise 3.2.16

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{23}{194} = 0.12$

Exercise 3.2.17

Find $P(\text{O})$.

Exercise 3.2.18

Find $P(\text{S})$.

Answer

$P(\text{S}) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise 3.2.19

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise 3.2.20

What is the probability of drawing a club in a standard deck of 52 cards?

Answer

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercise 3.2.21

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise 3.2.22

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Answer

$\frac{3}{6} = \frac{1}{2} = 0.5$

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.

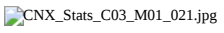


Figure 4.1.1.1.

Let $\text{B} =$ the event of landing on blue.

Let $\text{R} =$ the event of landing on red.

Let $\mathbf{G}$ = the event of landing on green.

Let $\mathbf{Y}$ = the event of landing on yellow.

Exercise 3.2.23

If you land on $\mathbf{Y}$, you get the biggest prize. Find $P(\mathbf{Y})$.

Exercise 3.2.24

If you land on red, you don't get a prize. What is $P(\mathbf{R})$?

Answer

$\mathbf{P}(R) = \frac{4}{8} = 0.5$

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 $\mathbf{I}$ = the event that a player in an infielder.

Let $\mathbf{O}$ = the event that a player is an outfielder.

Let $\mathbf{H}$ = the event that a player is a great hitter.

Let $\mathbf{N}$ = the event that a player is not a great hitter.

Exercise 3.2.25

Write the symbols for the probability that a player is not an outfielder.

Exercise 3.2.26

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Answer

$P(\mathbf{O}\ \mathbf{OR}\ \mathbf{H})$

Exercise 3.2.27

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise 3.2.28

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Answer

$P(\mathbf{H}|\mathbf{I})$

Exercise 3.2.29

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise 3.2.30

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Answer

$P(\text{N}|\text{O})$

Exercise 3.2.31

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise 3.2.32

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Answer

$P(\text{I OR N})$

Exercise 3.2.33

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise 3.2.34

Write the symbols for the probability that a player is an infielder.

Answer

$P(\text{I})$

Exercise 3.2.35

What is the word for the set of all possible outcomes?

Exercise 3.2.36

What is conditional probability?

Answer

The likelihood that an event will occur given that another event has already occurred.

Exercise 3.2.37

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 $\text{F} =$ event that book is fiction

Let $\text{N} =$ event that book is nonfiction

What is the sample space?

Exercise 3.2.38

What is the sum of the probabilities of an event and its complement?

Answer

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let $\text{E} =$ the event that it lands on an even number. Let $\text{M} =$ the event that it lands on a multiple of three.

Exercise 3.2.39

What does $P(\mathrm{E}|\mathrm{M})$ mean in words?

Exercise 3.2.40

What does $P(\mathrm{E\ OR\ M})$ mean in words?

Answer

the probability of landing on an even number or a multiple of three

This page titled 4.1.1: Terminology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.1.2: Independent and Mutually Exclusive Events

Independent and *mutually exclusive* do not mean the same thing.

Independent Events

Two events are independent if the following are true:

- $P(\text{A}|\text{B}) = P(\text{A})$
- $P(\text{B}|\text{A}) = P(\text{B})$
- $P(\text{A AND B}) = P(\text{A})P(\text{B})$

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 a population

Sampling may be done with replacement or without replacement (Figure **4.1.2.1**):

- **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*.

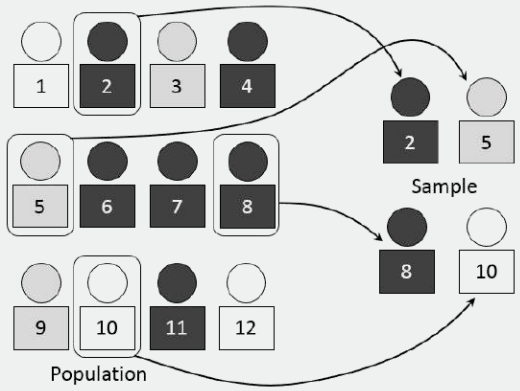


Figure **4.1.2.1**: A visual representation of the sampling process. If the sample items are replaced after each sampling event, then this is "sampling with replacement" if not, then it is "sampling without replacement". (CC BY-SA 4.0; Dan Kernler).

If it is not known whether **A** and **B** are independent or dependent, assume they are dependent until you can show otherwise.

Example 4.1.2.1: Sampling with and without replacement

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.

? Exercise 4.1.2.1

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?

Answer a

With replacement

Answer b

No

✓ Example 4.1.2.2

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, 1D, 1C, QD**.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are **KH, 7D, 6D, KH**.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Answer a

Without replacement

Answer b

With replacement

? Exercise 4.1.2.2

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, 1D, 1C, QD**
b. **KH, 7D, 6D, KH**
c. **QS, 7D, 6D, KS**

Answer - without replacement

a. Possible; b. Impossible, c. Possible

Answer - with replacement

a. Possible; c. Possible, c. Possible

Mutually Exclusive Events

A and B are mutually exclusive events if they **cannot** occur at the same time. This means that A and B do not share any outcomes and $P(\text{A AND B}) = 0$.

For example, suppose the sample space

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Let $\text{A} = \{1, 2, 3, 4, 5\}, \text{B} = \{4, 5, 6, 7, 8\}$, and $\text{C} = \{7, 9\}$. A AND B $= \{4, 5\}$.

$$P(\text{A AND B}) = \frac{2}{10}$$

and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so $P(\text{A AND C}) = 0$. 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 4.1.2.3

Flip two fair coins.

The sample space is $\{HH, HT, TH, TT\}$ where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (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 $\text{B} = \text{A}'$. Also, $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$.
- The probabilities for A and for B are $P(\text{A}) = \frac{3}{4}$ and $P(\text{B}) = \frac{1}{4}$.
- Let C = the event of getting all heads. $\text{C} = \{HH\}$. Since $\text{B} = \{TT\}$, $P(\text{B AND C}) = 0$. B and Care 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. $\text{D} = \{TT\}$. $P(\text{D}) = \frac{1}{4}$
- 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.) $\text{E} = \{HT, HH\}$. $P(\text{E}) = \frac{2}{4}$
- 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. $\text{F} = \{HT, TH, TT\}$. $P(\text{F}) = \frac{3}{4}$

Exercise 4.1.2.3

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Answer

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is $\{BB, BR, RB, RR\}$.

Event A = Getting at least one black card $= \{BB, BR, RB\}$

$P(\text{A}) = \frac{3}{4} = 0.75$

Example 4.1.2.4

Flip two fair coins. Find the probabilities of the events.

a. Let $\mathbf{F}$ = the event of getting at most one tail (zero or one tail).
b. Let $\mathbf{G}$ = the event of getting two faces that are the same.
c. Let $\mathbf{H}$ = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are $\mathbf{F}$ and $\mathbf{G}$ mutually exclusive?
e. Let $\mathbf{J}$ = the event of getting all tails. Are $\mathbf{J}$ and $\mathbf{H}$ mutually exclusive?

Solution

Look at the sample space in Example 4.1.2.3.

1. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(\mathbf{F}) = \frac{3}{4}$
2. Two faces are the same if HH or TT show up. $P(\mathbf{G}) = \frac{2}{4}$
3. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(\mathbf{H}) = \frac{2}{4}$
4. $\mathbf{F}$ and $\mathbf{G}$ share HH so $P(\mathbf{F}\text{ AND }\mathbf{G})$ is not equal to zero (0). $\mathbf{F}$ and $\mathbf{G}$ are not mutually exclusive.
5. Getting all tails occurs when tails shows up on both coins (TT). $\mathbf{H}$'s outcomes are HH and HT.

$\mathbf{J}$ and $\mathbf{H}$ have nothing in common so $P(\mathbf{J}\text{ AND }\mathbf{H}) = 0$. $\mathbf{J}$ and $\mathbf{H}$ are mutually exclusive.

Exercise 4.1.2.4

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 $\mathbf{F}$ = the event of getting the white ball twice.
b. Let $\mathbf{G}$ = the event of getting two balls of different colors.
c. Let $\mathbf{H}$ = the event of getting white on the first pick.
d. Are $\mathbf{F}$ and $\mathbf{G}$ mutually exclusive?
e. Are $\mathbf{G}$ and $\mathbf{H}$ mutually exclusive?

Answer

a. $P(\mathbf{F}) = \frac{1}{4}$
b. $P(\mathbf{G}) = \frac{1}{2}$
c. $P(\mathbf{H}) = \frac{1}{2}$
d. Yes
e. No

Example 4.1.2.5

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event $\mathbf{A}$ = a face is odd. Then $\mathbf{A} = \{1, 3, 5\}$. Let event $\mathbf{B}$ = a face is even. Then $\mathbf{B} = \{2, 4, 6\}$.

- Find the complement of $\mathbf{A}$, $\mathbf{A}'$. The complement of $\mathbf{A}$, $\mathbf{A}'$, is $\mathbf{B}$ because $\mathbf{A}$ and $\mathbf{B}$ together make up the sample space. $P(\mathbf{A}) + P(\mathbf{B}) = P(\mathbf{A}) + P(\mathbf{A}') = 1$. Also, $P(\mathbf{A}) = \frac{3}{6}$ and $P(\mathbf{B}) = \frac{3}{6}$.
- Let event $\mathbf{C}$ = odd faces larger than two. Then $\mathbf{C} = \{3, 5\}$. Let event $\mathbf{D}$ = all even faces smaller than five. Then $\mathbf{D} = \{2, 4\}$. $P(\mathbf{C}\text{ AND }\mathbf{D}) = 0$ because you cannot have an odd and even face at the same time. Therefore, $\mathbf{C}$ and $\mathbf{D}$ are mutually exclusive events.
- Let event $\mathbf{E}$ = all faces less than five. $\mathbf{E} = \{1, 2, 3, 4\}$.

Are $\mathbf{C}$ and $\mathbf{E}$ mutually exclusive events? (Answer yes or no.) Why or why not?

Answer

No. $C = \{3, 5\}$ and $E = \{1, 2, 3, 4\}$. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = \frac{2}{3}$.

? Exercise 4.1.2.5

Let event A = learning Spanish. Let event B = learning German. Then $A \text{ AND } B$ = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(A|B) = P(A)$
- $P(B|A)$
- $P(A \text{ AND } B) = P(A)P(B)$

Answer

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{0.08}{0.2} = 0.4 = P(A) \quad (4.1.2.1)$$

The events are independent because $P(A|B) = P(A)$.

✓ Example 4.1.2.6

Let event G = taking a math class. Let event H = taking a science class. Then, $G \text{ AND } H$ = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

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. a. Show that $P(G|H) = P(G)$.
b. b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution

a. $P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$
b. $P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ AND } H)$

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 he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

? Exercise 4.1.2.6

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 $P(\text{G AND O})$?

Answer

Event G and O $= \{G1, G3\}$

$P(\text{G and O}) = \frac{2}{10} = 0.2$

Example 4.1.2.7

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(\text{C}) = 0.75$, $P(\text{D}) = 0.3$, $P(\text{C|D}) = 0.75$ and $P(\text{C AND D}) = 0.225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(\text{D|C})$?

Solution

a. Yes, because $P(\text{C|D}) = P(\text{C})$.
b. No, because $P(\text{C AND D})$ is not equal to zero.
c. $P(\text{D|C}) = \frac{P(\text{C AND D})}{P(\text{C})} = \frac{0.225}{0.75} = 0.3$

Exercise 4.1.2.7

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 $P(\text{B}) = 0.40$, $P(\text{D}) = 0.30$ and $P(\text{B AND D}) = 0.20$.

a. Find $P(\text{B|D})$.
b. Find $P(\text{D|B})$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Answer

a. $P(\text{B|D}) = 0.6667$
b. $P(\text{D|B}) = 0.5$
c. No
d. No

Example 4.1.2.8

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.

- $P(\text{R}) = \frac{3}{8}$. $P(\text{B}) = \frac{5}{8}$. $P(\text{R AND B}) = 0$. (You cannot draw one card that is both red and blue.)

- $P(\mathrm{E}) = \frac{3}{8}$. (There are three even-numbered cards, $R2, B2$, and $B4$.)
- $P(\mathrm{E}|\mathrm{B}) = \frac{2}{5}$. (There are five blue cards: $B1, B2, B3, B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(\mathrm{B}|\mathrm{E}) = \frac{2}{3}$. (There are three even-numbered cards: $R2, B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(\mathrm{R\ AND\ B}) = 0$.
- Let G = card with a number greater than 3. $\mathrm{G} = \{B4, B5\}$. $P(\mathrm{G}) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $\mathrm{H} = \{B1, B2, B3, B4\}$. $P(\mathrm{G}|\mathrm{H}) = \frac{1}{4}$. (The only card in H that has a number greater than three is B4.) Since $\frac{2}{8} = \frac{1}{4}$, $P(\mathrm{G}) = P(\mathrm{G}|\mathrm{H})$, which means that G and H are independent.

? Exercise 4.1.2.8

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?

Answer

- $P(\mathrm{B}|\mathrm{A}) = 0.67$
- $P(\mathrm{B}) = 0.25$

So $P(\mathrm{B})$ does not equal $P(\mathrm{B}|\mathrm{A})$ which means that $\mathrm{B and A}$ are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\mathrm{B\ AND\ A}) = 0.20$, not 0.

✓ Example 4.1.2.9

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:
- $P(\mathrm{F}) = 0.60$; $P(\mathrm{L}) = 0.50$
- $P(\mathrm{F\ AND\ L}) = 0.45$
- $P(\mathrm{L}|\mathrm{F}) = 0.75$

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 $P(\mathrm{F}|\mathrm{L})$ yet, so you cannot use the second condition.

Solution 1

Check whether $P(\mathrm{F\ AND\ L}) = P(\mathrm{F})P(\mathrm{L})$. We are given that $P(\mathrm{F\ AND\ L}) = 0.45$, but $P(\mathrm{F})P(\mathrm{L}) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(\mathrm{F\ AND\ L})$ does not equal $P(\mathrm{F})P(\mathrm{L})$.

Solution 2

Check whether $P(\mathrm{L}|\mathrm{F})$ equals $P(\mathrm{L})$. We are given that $P(\mathrm{L}|\mathrm{F}) = 0.75$, but $P(\mathrm{L}) = 0.50$; 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 4.1.2.9

Mark is deciding which route to take to work. His choices are $\mathbf{I = \text{the Interstate}}$ and $\mathbf{F = \text{Fifth Street}}$

- $P(\mathbf{I}) = 0.44$ and $P(\mathbf{F}) = 0.55$
- $P(\mathbf{I \text{ AND } F}) = 0$ because Mark will take only one route to work.

What is the probability of $P(\mathbf{I \text{ OR } F})$?

Answer

Because $P(\mathbf{I \text{ AND } F}) = 0$,

$P(\mathbf{I \text{ OR } F}) = P(\mathbf{I}) + P(\mathbf{F}) - P(\mathbf{I \text{ AND } F}) = 0.44 + 0.56 - 0 = 1$

Example 4.1.2.10

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 in three 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 $\mathbf{A}$ = heads ($\mathbf{H}$) on the coin followed by an even number (2, 4, 6) on the die.
$\mathbf{A}$ = {_____}. Find $P(\mathbf{A})$.
f. Event $\mathbf{B}$ = heads on the coin followed by a three on the die. $\mathbf{B}$ = {_____}. Find $P(\mathbf{B})$.
g. Are $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive? (Hint: What is $P(\mathbf{A \text{ AND } B})$? If $P(\mathbf{A \text{ AND } B}) = 0$, then $\mathbf{A}$ and $\mathbf{B}$ are mutually exclusive.)
h. Are $\mathbf{A}$ and $\mathbf{B}$ independent? (Hint: Is $P(\mathbf{A \text{ AND } B}) = P(\mathbf{A})P(\mathbf{B})$? If $P(\mathbf{A \text{ AND } B}) = P(\mathbf{A})P(\mathbf{B})$, then $\mathbf{A}$ and $\mathbf{B}$ are independent. If not, then they are dependent).

Solution

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. $\mathbf{A} = \{H2, H4, H6\}$; $P(\mathbf{A}) = \frac{3}{12}$
f. $\mathbf{B} = \{H3\}$; $P(\mathbf{B}) = \frac{1}{12}$
g. Yes, because $P(\mathbf{A \text{ AND } B}) = 0$
h. $P(\mathbf{A \text{ AND } B}) = 0$. $P(\mathbf{A})P(\mathbf{B}) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(\mathbf{A \text{ AND } B})$ does not equal $P(\mathbf{A})P(\mathbf{B})$, so $\mathbf{A}$ and $\mathbf{B}$ are dependent.

Exercise 4.1.2.10

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.

a. Compute $P(\mathbf{T})$.
b. Compute $P(\mathbf{T}|\mathbf{F})$.
c. Are **T** and **F** independent?.
d. Are **F** and **S** mutually exclusive?

e. Are F and S independent?

Answer

a. $P(\text{T}) = \frac{1}{4}$
b. $P(\text{T}|\text{F}) = \frac{1}{2}$
c. No
d. No
e. Yes

References

1. 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).
2. Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Review

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. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

- If A and B are independent, $P(\text{A AND B}) = P(\text{A})P(\text{B}), P(\text{A}|\text{B}) = P(\text{A})$ and $P(\text{B}|\text{A}) = P(\text{B})$.
- If A and B are mutually exclusive, $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) and P(\text{A AND B}) = 0$.

? Exercise 4.1.2.11

E and F are mutually exclusive events. $P(\text{E}) = 0.4$; $P(\text{F}) = 0.5$. Find $P(\text{E}|\text{F})$.

? Exercise 4.1.2.12

J and K are independent events. $P(\text{J}|\text{K}) = 0.3$. Find $P(\text{J})$.

Answer

$P(\text{J}) = 0.3$

? Exercise 4.1.2.13

U and V are mutually exclusive events. $P(\text{U}) = 0.26$; $P(\text{V}) = 0.37$. Find:

a. $P(\text{U AND V}) =$
b. $P(\text{U}|\text{V}) =$
c. $P(\text{U OR V}) =$

? Exercise 4.1.2.14

Q and R are independent events. $P(\text{Q}) = 0.4$ and $P(\text{Q AND R}) = 0.1$. Find $P(\text{R})$.

Answer

$P(\text{Q AND R}) = P(\text{Q})P(\text{R})$

$0.1 = (0.4)P(\text{R})$

$P(\text{R}) = 0.25$

Bringing It Together

Exercise 4.1.2.16

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into Table.

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.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt\#1–33} | \leq 210 \text{ pounds})$?

Exercise 4.1.2.17

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

a. $P(\text{C}) =$ _____
b. $P(\text{P}|\text{C}) =$ _____
c. $P(\text{P}|\text{C'}) =$ _____
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Answer

a. $P(\text{C}) = 0.4567$
b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise 4.1.2.18

Given events G and H : $P(\text{G}) = 0.43$; $P(\text{H}) = 0.26$; $P(\text{H AND G}) = 0.14$

a. Find $P(\text{H OR G})$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

Exercise 4.1.2.19

Given events J and K : $P(\text{J}) = 0.18$; $P(\text{K}) = 0.37$; $P(\text{J OR K}) = 0.45$

a. Find $P(\text{J AND K})$.
b. Find the probability of the complement of event (J AND K).

c. Find the probability of the complement of event (J AND K).

Answer

a. $P(\text{J OR K}) = P(\text{J}) + P(\text{K}) - P(\text{J AND K}); 0.45 = 0.18 + 0.37 - P(\text{J AND K})$; solve to find $P(\text{J AND K}) = 0.10$
b. $P(\text{NOT (J AND K)}) = 1 - P(\text{J AND K}) = 1 - 0.10 = 0.90$
c. $P(\text{NOT (J OR K)}) = 1 - P(\text{J OR K}) = 1 - 0.45 = 0.55$

Glossary

Dependent Events

If two events are NOT independent, then we say that they are dependent.

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 Conditional Probability of One Event Given Another Event

$P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The OR of Two Events

An outcome is in the event A OR B if the outcome is in A, is in B, or is in both A and B.

This page titled 4.1.2: Independent and Mutually Exclusive Events is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.2: Addition and Multiplication Rule 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:

$$P(A \text{ AND } B) = P(B)P(A|B) \tag{4.2.1}$$

This rule may also be written as:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

(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

$$P(A|B) = P(A).$$

and Equation 4.2.1 becomes

$$P(A \text{ AND } B) = P(A)P(B).$$

The Addition Rule

If A and B are defined on a sample space, then:

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) \tag{4.2.2}$$

If A and B are **mutually exclusive**, then

$$P(A \text{ AND } B) = 0.$$

and Equation 4.2.2 becomes

$$P(A \text{ OR } B) = P(A) + P(B).$$

Example 4.2.1

Klaus is trying to choose where to go on vacation. His two choices are: $\mathbf{A} = \textbf{New Zealand}$ and $\mathbf{B} = \textbf{Alaska}$.

- Klaus can only afford one vacation. The probability that he chooses $\mathbf{A}$ is $P(\mathbf{A}) = 0.6$ and the probability that he chooses $\mathbf{B}$ is $P(\mathbf{B}) = 0.35$.
- $P(\mathbf{A} \text{ AND } \mathbf{B}) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(\mathbf{A} \text{ OR } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

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. $\mathbf{A}$ = the event Carlos is successful on his first attempt. $P(\mathbf{A}) = 0.65$. $\mathbf{B}$ = the event Carlos is successful on his second attempt. $P(\mathbf{B}) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** 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 $\mathbf{A}$ and $\mathbf{B}$ independent?
d. Are $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive?

Solutions

a. The problem is asking you to find $P(\text{A AND B}) = P(\text{B AND A})$. Since $P(\text{B}|\text{A}) = 0.90 : P(\text{B AND A}) = P(\text{B}|\text{A})P(\text{A}) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

b. The problem is asking you to find $P(\text{A OR B})$.

$$P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B}) = 0.65 + 0.65 - 0.585 = 0.715 \tag{4.2.3}$$

Carlos makes either the first goal or the second goal with probability 0.715.

c. No, they are not, because $P(\text{B AND A}) = 0.585$.

$$P(\text{B})P(\text{A}) = (0.65)(0.65) = 0.423 \tag{4.2.4}$$

$$0.423 \neq 0.585 = P(\text{B AND A}) \tag{4.2.5}$$

So, $P(\text{B AND A})$ is **not** equal to $P(\text{B})P(\text{A})$.

d. No, they are not because $P(\text{A and B}) = 0.585$.

To be mutually exclusive, $P(\text{A AND B})$ must equal zero.

Exercise 4.2.1

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(\text{C}) = 0.75$. D = the event Helen makes the second shot. $P(\text{D}) = 0.75$. 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?

Answer

$$P(\text{D}|\text{C}) = 0.85 \tag{4.2.6}$$

$$P(\text{C AND D}) = P(\text{D AND C}) \tag{4.2.7}$$

$$P(\text{D AND C}) = P(\text{D}|\text{C})P(\text{C}) = (0.85)(0.75) = 0.6375 \tag{4.2.8}$$

Helen makes the first and second free throws with probability 0.6375.

Example 4.2.2

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?
b. What is the probability that the member practices four times a week?
c. What is the probability that the member is an advanced swimmer and practices four times a week?
d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?
e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Answer

a. $\frac{28}{150}$
b. $\frac{80}{150}$
c. $\frac{40}{150}$
d. $P(\text{advanced AND intermediate}) = 0$, 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.

$$P(\text{novice AND practices four times per week}) = 0.0667 \tag{4.2.9}$$

$$P(\text{novice})P(\text{practices four times per week}) = 0.0996 \tag{4.2.10}$$

$$0.0667 \neq 0.0996 \tag{4.2.11}$$

Exercise 4.2.2

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?

Answer

$$P = \frac{200 - 140 - 40}{200} = \frac{20}{200} = 0.1 \tag{4.2.12}$$

Example 4.2.3

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 GIVEN that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, $\mathrm{M|S}$ = math given speech

a. What is the probability that Felicity enrolls in math and speech?
   Find $P(\mathrm{M\ AND\ S}) = P(\mathrm{M|S})P(\mathrm{S})$.
b. What is the probability that Felicity enrolls in math or speech classes?
   Find $P(\mathrm{M\ OR\ S}) = P(\mathrm{M}) + P(\mathrm{S}) - P(\mathrm{M\ AND\ S})$.
c. Are M and S independent? Is $P(\mathrm{M|S}) = P(\mathrm{M})$?
d. Are M and S mutually exclusive? Is $P(\mathrm{M\ AND\ S}) = 0$?

Answer

a. 0.1625, b. 0.6875, c. No, d. No

Exercise 4.2.3

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 $P(\mathrm{B}) = 0.40, P(\mathrm{D}) = 0.30$ and $P(\mathrm{D|B}) = 0.5$.

a. Find $P(\mathrm{B\ AND\ D})$.
b. Find $P(\mathrm{B\ OR\ D})$.

Answer

a. $P(\mathrm{B\ AND\ D}) = P(\mathrm{D|B})P(\mathrm{B}) = (0.5)(0.4) = 0.20$.
b. $P(\mathrm{B\ OR\ D}) = P(\mathrm{B}) + P(\mathrm{D}) - P(\mathrm{B\ AND\ D}) = 0.40 + 0.30 - 0.20 = 0.50$

Example 4.2.4

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?

Answers

a. $P(\text{B}) = 0.143; P(\text{N}) = 0.85$
b. $P(\text{N}|\text{B}) = 0.02$
c. $P(\text{B AND N}) = P(\text{B})P(\text{N}|\text{B}) = (0.143)(0.02) = 0.0029$
d. $P(\text{B OR N}) = P(\text{B}) + P(\text{N}) - P(\text{B AND N}) = 0.143 + 0.85 - 0.0029 = 0.9901$
e. No. $P(\text{N}) = 0.85; P(\text{N}|\text{B}) = 0.02$. So, $P(\text{N}|\text{B})$ does not equal $P(\text{N})$.
f. No. $P(\text{B AND N}) = 0.0029$. For B and N to be mutually exclusive, $P(\text{B AND N})$ must be zero

Exercise 4.2.4

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?

Answer

Let $\text{A} =$ student is a senior going to college.

Let $\text{B} =$ student plays sports.

$P(\text{B}) = \frac{140}{200}$

$P(\text{B}|\text{A}) = \frac{50}{140}$

$P(\text{A AND B}) = P(\text{B}|\text{A})P(\text{A})$

$P(\text{A AND B}) = (\frac{140}{200})(\frac{50}{140}) = \frac{1}{4}$

Example 4.2.5

Refer to the information in Example 4.2.4. $\text{P} =$ tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(\text{P}|\text{B}) = 1 - P(\text{N}|\text{B})$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(\text{B AND P}) = P(\text{P}|\text{B})P(\text{B})$.
c. What is the probability that a woman does not develop breast cancer. Find $P(\text{B}') = 1 - P(\text{B})$.
d. What is the probability that a woman tests positive for breast cancer. Find $P(\text{P}) = 1 - P(\text{N})$.

Answer

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Exercise 4.2.5

A student goes to the library. Let events $\text{B} =$ the student checks out a book and $\text{D} =$ the student checks out a DVD. Suppose that $P(\text{B}) = 0.40, P(\text{D}) = 0.30$ and $P(\text{D}|\text{B}) = 0.5$.

a. Find $P(\text{B}')$.
b. Find $P(\text{D AND B})$.
c. Find $P(\text{B}|\text{D})$.
d. Find $P(\text{D AND B}')$.
e. Find $P(\text{D}|\text{B}')$.

Answer

a. $P(\text{B}') = 0.60$
b. $P(\text{D AND B}) = P(\text{D}|\text{B})P(\text{B}) = 0.20$
c. $P(\text{B}|\text{D}) = \frac{P(\text{B AND D})}{P(\text{D})} = \frac{(0.20)}{(0.30)} = 0.66$
d. $P(\text{D AND B}') = P(\text{D}) - P(\text{D AND B}) = 0.30 - 0.20 = 0.10$
e. $P(\text{D}|\text{B}') = P(\text{D AND B}')P(\text{B}') = (P(\text{D}) - P(\text{D AND B}))(0.60) = (0.10)(0.60) = 0.06$

References

1. DiCamillo, Mark, Mervin Field. "The File Poll." Field Research Corporation. Available online at www.field.com/fieldpollonline...rs/Rls2443.pdf (accessed May 2, 2013).
2. Rider, David, "Ford support plummeting, poll suggests," The Star, September 14, 2011. Available online at www.thestar.com/news/gta/2011..._suggests.html (accessed May 2, 2013).
3. "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).
4. "Roulette." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Roulette (accessed May 2, 2013).
5. 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).
6. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).
7. Data from U.S. Census Bureau.
8. Data from the Wall Street Journal.
9. Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at www.ropercenter.uconn.edu/ (accessed May 2, 2013).
10. Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Review

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. 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 to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(\text{A AND B}) = P(\text{A}|\text{B})P(\text{B})$

The addition rule: $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B})$

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.

Exercise 4.2.5

Find $P(C)$.

Exercise 4.2.6

Find $P(L)$.

Answer

0.376

Exercise 4.2.7

Find $P(C|L)$.

Exercise 4.2.8

In words, what is $C|L$?

Answer

$C|L$ 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.

Exercise 4.2.9

Find $P(L \text{ AND } C)$

Exercise 4.2.10

In words, what is $L \text{ AND } C$?

Answer

$L \text{ AND } C$ 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.

Exercise 4.2.11

Are L and C independent events? Show why or why not.

Exercise 4.2.12

Find $P(L \text{ OR } C)$.

Answer

0.6492

Exercise 4.2.13

In words, what is $L \text{ OR } C$?

? Exercise 4.2.14

Are $\mathbf{L}$ and $\mathbf{C}$ mutually exclusive events? Show why or why not.

Answer

No, because $P(\mathbf{L} \text{ AND } \mathbf{C})$ does not equal 0.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events $\mathbf{A}$ and $\mathbf{B}$ are independent if one of the following is true:

1. $P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A})$
2. $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$
3. $P(\mathbf{A} \text{ AND } \mathbf{B}) = P(\mathbf{A})P(\mathbf{B})$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events $\mathbf{A}$ and $\mathbf{B}$ are mutually exclusive, then $P(\mathbf{A} \text{ AND } \mathbf{B}) = 0$.

This page titled 4.2: Addition and Multiplication Rule of Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **3.4: Two Basic Rules of Probability** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

4.3: Conditional Probability using 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 4.3.1

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

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	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 $P(\text{Person is a cell phone user})$.
b. Find $P(\text{person had no violation in the last year})$.
c. Find $P(\text{Person had no violation in the last year AND was a cell phone user})$.
d. Find $P(\text{Person is a cell phone user OR person had no violation in the last year})$.
e. Find $P(\text{Person is a cell phone user GIVEN person had a violation in the last year})$.
f. Find $P(\text{Person had no violation last year GIVEN person was not a cell phone user})$

Answer

a. $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$
b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$
c. $\frac{280}{755}$
d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$
e. $\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)
f. $\frac{405}{450}$ (The sample space is reduced to the number of persons who were not cell phone users.)

Exercise 4.3.1

Table 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

a. What is $P(\text{athlete stretches before exercising})$?
b. What is $P(\text{athlete stretches before exercising}|\text{no injury in the last year})$?

Answer

a. $P(\text{athlete stretches before exercising}) = \frac{350}{800} = 0.4375$

b. $P(\text{athlete stretches before exercising}|\text{no injury in the last year}) = \frac{295}{514} = 0.5739$

Example 4.3.2

Table shows a random sample of 100 hikers and the areas of hiking they prefer.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

a. Complete the table.

b. Are the events "being female" and "preferring the coastline" independent events? Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$
3. Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let $\mathbf{M} =$ being male, and let $\mathbf{L} =$ 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?

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let $\mathbf{F} =$ being female, and let $\mathbf{P} =$ prefers mountain peaks.

1. Find $P(\mathbf{F})$.
2. Find $P(\mathbf{P})$.
3. Find $P(\mathbf{F} \text{ AND } \mathbf{P})$.
4. Find $P(\mathbf{F} \text{ OR } \mathbf{P})$.

Answers

a.

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.

$$P(\mathbf{F} \text{ AND } \mathbf{C}) = \frac{18}{100} = 0.18$$

$$P(\text{F})P(\text{C}) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(\text{F AND C}) \neq P(\text{F})P(\text{C})$, so the events F and C are not independent.

c.

1. The word 'given' tells you that this is a conditional.
2. $P(\text{M}|\text{L}) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d.

a. Find $P(\text{F})$.
b. Find $P(\text{P})$.
c. Find $P(\text{F AND P})$.
d. Find $P(\text{F OR P})$.

d.

1. $P(\text{F}) = \frac{45}{100}$
2. $P(\text{P}) = \frac{25}{100}$
3. $P(\text{F AND P}) = \frac{11}{100}$
4. $P(\text{F OR P}) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Exercise 4.3.2

Table shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

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?

Answer

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$
$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$
$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example 4.3.3

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

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 $P(\text{Door One AND Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Not Caught})$

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.
b. What is the probability that Alissa does not catch Muddy?
c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution

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. $\frac{41}{60}$
c. $\frac{9}{19}$

Example 4.3.4

Table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

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

a. Find $P(\text{2009 AND Robbery})$.
b. Find $P(\text{2010 AND Burglary})$.
c. Find $P(\text{2010 OR Burglary})$.
d. Find $P(\text{2011}|\text{Rape})$.
e. Find $P(\text{Vehicle}|\text{2008})$.

Answer

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Exercise 4.3.3

Table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Answer

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. $P(\text{Tall}) = \frac{50}{205} = 0.244$
c. $P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$
d. $P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$
e. $P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$
f. $P(\text{Tall AND Underweight}) = \frac{12}{205} = 0.0585$
g. No. $P(\text{Tall})$ does not equal $P(\text{Tall}|\text{Obese})$.

References

1. “Blood Types.” American Red Cross, 2013. Available online at www.redcrossblood.org/learn-a...od/blood-types (accessed May 3, 2013).
2. Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.
3. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).
4. 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).

5. “Human Blood Types.” Unite Blood Services, 2011. Available online at www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).
6. Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at www.ehow.com/facts_5552003_st...ive-blood.html (accessed May 2, 2013).
7. “United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Use the following information to answer the next four exercises. Table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

? Exercise 3.5.4

Find P(musician is a female).

? Exercise 3.5.5

Find $P(\text{musician is a male AND had private instruction})$.

Answer

$$P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$$

? Exercise 3.5.6

Find P(musician is a female OR is self taught).

? Exercise 3.5.7

Are the events “being a female musician” and “learning music in school” mutually exclusive events?

Answer

The events are not mutually exclusive. It is possible to be a female musician who learned music in school.

Bringing it Together

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.

Exercise 3.5.8

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Exercise 3.5.9

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Answer

$\frac{35,065}{100,450}$

Exercise 3.5.10

Find the probability that the person was Latino.

Exercise 3.5.11

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.

Answer

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 $\frac{4,715}{100,450}$.

Exercise 3.5.12

In words, explain what it means to pick one person from the study who is "Japanese American **OR** smokes 21 to 30 cigarettes per day." Also, find the probability.

Exercise 3.5.13

In words, explain what it means to pick one person from the study who is "Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

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 $\frac{4,715}{15,273}$.

? Exercise 3.5.14

Prove that smoking level/day and ethnicity are dependent events.

Glossary

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.

This page titled 4.3: Conditional Probability using Contingency Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.E: Probability Topics (Exericses)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

3.1: Introduction

3.2: Terminology

Q 3.2.1

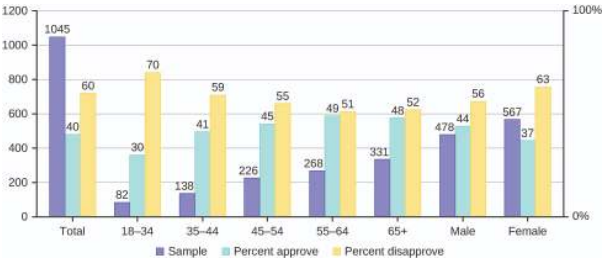


Figure 3.2.3.2.11.

The graph in Figure 3.2.1 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.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

Q 3.2.2

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.

S 3.2.2

a. 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%
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

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.

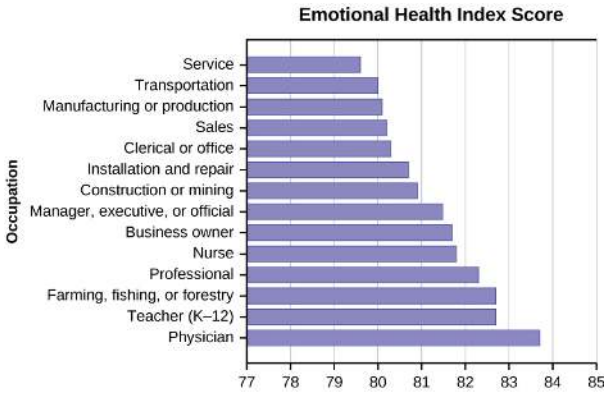


Figure 3.3.1.

Q 3.3.1

Find the probability that an Emotional Health Index Score is 82.7.

Q 3.3.2

Find the probability that an Emotional Health Index Score is 81.0.

S 3.3.2

0

Q 3.3.3

Find the probability that an Emotional Health Index Score is more than 81?

Q 3.3.4

Find the probability that an Emotional Health Index Score is between 80.5 and 82?

S 3.3.4

0.3571

Q 3.3.5

If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

Q 3.3.6

What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

S 3.3.6

0.2142

Q 3.3.7

What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

Q 3.3.8

What occupation has the highest emotional index score?

S 3.3.8

Physician (83.7)

Q 3.3.9

What occupation has the lowest emotional index score?

Q 3.3.10

What is the range of the data?

S 3.3.10

83.7 − 79.6 = 4.1

Q 3.3.11

Compute the average EHIS.

Q 3.3.12

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?

S 3.3.12

$P(\text{Occupation} < 81.3) = 0.5$

3.4: Two Basic Rules of Probability

Q 3.4.1

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:

- $\mathbf{C}$ = California registered voters who support same-sex marriage.
- $\mathbf{B}$ = 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
- $\mathbf{A}$ = California registered voters who are 18 to 39 years old.

a. Find $P(\mathbf{C})$.
b. Find $P(\mathbf{B})$.
c. Find $P(\mathbf{C}|\mathbf{A})$.
d. Find $P(\mathbf{B}|\mathbf{C})$.
e. In words, what is $\mathbf{C}|\mathbf{A}$?
f. In words, what is $\mathbf{B}|\mathbf{C}$?
g. Find $P(\mathbf{C}\ \mathbf{AND}\ \mathbf{B})$.
h. In words, what is $\mathbf{C}\ \mathbf{AND}\ \mathbf{B}$?
i. Find $P(\mathbf{C}\ \mathbf{OR}\ \mathbf{B})$.
j. Are $\mathbf{C}$ and $\mathbf{B}$ mutually exclusive events? Show why or why not.

Q 3.4.2

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?

S 3.4.2

a. The Forum Research surveyed 1,046 Torontonians.
b. 58%
c. 42% of 1,046 = 439 (rounding to the nearest integer)
d. 0.57
e. 0.60.

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.

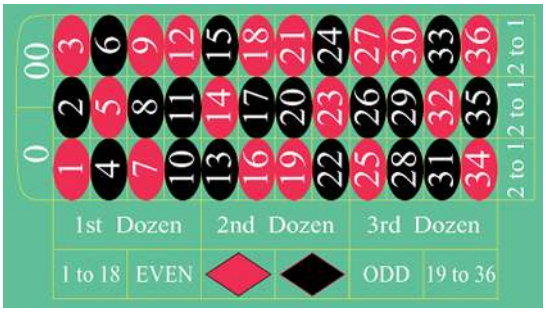


Figure 3.4.1

Q 3.4.3

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find $P(\text{red})$.
c. You bet on -1st 12- (1st Dozen). Find $P(\text{-1st 12-})$.
d. You bet on an even number. Find $P(\text{even number})$.
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

Q 3.4.4

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

S 3.4.4

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. $P(\text{Betting on 0-00-1-2-3}) = \frac{5}{38}$

https://stats.libretexts.org/@go/page/48359

7. $P(\text{Betting on 0-1-2; or 0-00-2; or 00-2-3}) = \frac{3}{38}$

Q 3.4.5

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)

Q 3.4.6

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.

- $\mathbf{G}$ = card drawn is green
- $\mathbf{E}$ = card drawn is even-numbered

a. List the sample space.
b. $P(\mathbf{G}) =$ _____
c. $P(\mathbf{G}|\mathbf{E}) =$ _____
d. $P(\mathbf{G \text{ AND } E}) =$ _____
e. $P(\mathbf{G \text{ OR } E}) =$ _____
f. Are $\mathbf{G}$ and $\mathbf{E}$ mutually exclusive? Justify your answer numerically.

S 3.4.6

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 $P(\mathbf{G \text{ AND } E})$ does not equal 0.

Q 3.4.7

Roll two fair dice. Each die has six faces.

a. List the sample space.
b. Let $\mathbf{A}$ be the event that either a three or four is rolled first, followed by an even number. Find $P(\mathbf{A})$.
c. Let B be the event that the sum of the two rolls is at most seven. Find $P(\mathbf{B})$.
d. In words, explain what "$P(\mathbf{A}|\mathbf{B})$" represents. Find $P(\mathbf{A}|\mathbf{B})$.
e. Are $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are $\mathbf{A}$ and $\mathbf{B}$ independent events? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.8

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let $\mathbf{A}$ be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(\mathbf{A})$.
c. Let $\mathbf{B}$ be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

d. Let **C** be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events **A** and **C** mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

S 3.4.9

The coin toss is independent of the card picked first.

a. $\{(G,H)(G,T)(B,H)(B,T)(R,H)(R,T)\}$
b. $P(\text{A}) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
c. 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). $P(\text{A AND B}) = 0$
d. 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). $P(\text{A AND C}) = P(\text{A}) = 320$

Q 3.4.10

An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let **A** be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(\text{A})$.
c. Let **B** be the event that the first and second tosses land on heads. Are the events **A** and **B** mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.11

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let **A** be the event that there are at least two tails. Find $P(\text{A})$.
c. Let **B** be the event that the first and second tosses land on heads. Are the events **A** and **B** mutually exclusive? Explain your answer in one to three complete sentences, including justification.

S 3.4.12

a. $S = (\text{HHH}), (\text{HHT}), (\text{HTH}), (\text{HTT}), (\text{THH}), (\text{THT}), (\text{TTH}), (\text{TTT})$
b. $\frac{4}{8}$
c. Yes, because if **A** has occurred, it is impossible to obtain two tails. In other words, $P(\text{A AND B}) = 0$.

Q 3.4.13

Consider the following scenario:

Let $P(\text{C}) = 0.4$.

Let $P(\text{D}) = 0.5$.

Let $P(\text{C}|\text{D}) = 0.6$.

a. Find $P(CANDD)$.
b. Are **C** and **D** mutually exclusive? Why or why not?
c. Are **C** and **D** independent events? Why or why not?
d. Find $P(\text{C OR D})$.
e. Find $P(\text{D}|\text{C})$.

Q 3.4.14

Y and **Z** are independent events.

a. Rewrite the basic Addition Rule $P(\text{Y OR Z}) = P(\text{Y}) + P(\text{Z}) - P(\text{Y AND Z})$ using the information that **Y** and **Z** are independent events.
b. Use the rewritten rule to find $P(\text{Z})$ if $P(\text{Y OR Z}) = 0.71$ and $P(\text{Y}) = 0.42$.

S 3.4.14

a. If Y and Z are independent, then $P(\mathrm{Y\ AND\ Z}) = P(\mathrm{Y})P(\mathrm{Z})$, so $P(\mathrm{Y\ OR\ Z}) = P(\mathrm{Y}) + P(\mathrm{Z}) - P(\mathrm{Y})P(\mathrm{Z})$.
b. 0.5

Q 3.4.15

G and H are mutually exclusive events. $P(\mathrm{G}) = 0.5 P(\mathrm{H}) = 0.3$

a. Explain why the following statement MUST be false: $P(\mathrm{H}|\mathrm{G}) = 0.4$.
b. Find $P(\mathrm{H\ OR\ G})$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

Q 3.4.16

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; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. $P(\mathrm{E}') =$	i. 0.8043
b. $P(\mathrm{E}) =$	ii. 0.623
c. $P(\mathrm{S\ and\ E}') =$	iii. 0.1957
d. $P(\mathrm{S}|\mathrm{E}') =$	iv. 0.1219

S 3.4.16

a. iii
b. i
c. iv
d. ii

Q 3.4.17

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.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Q 3.4.18

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; O = other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.

c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

S 3.4.18

a. $P(\text{R}) = 0.44$
b. $P(\text{R}|\text{E}) = 0.56$
c. $P(\text{R}|\text{O}) = 0.31$
d. 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; $P(\text{R}|\text{E}) \neq P(\text{R})$.
e. 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; $P(\text{R}|\text{E}) > P(\text{R})$.

Q 3.4.19

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.

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?

a. Yes, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit is a double})$
c. No, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
d. Yes, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) = P(\text{hit is a double})$

Q 3.4.29

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 blood type. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

S 3.4.30

a. $P(\text{type O OR Rh-}) = P(\text{type O}) + P(\text{Rh-}) - P(\text{type O AND Rh-})$

$0.52 = 0.43 + 0.15 - P(\text{type O AND Rh-})$; solve to find $P(\text{type O AND Rh-}) = 0.06$

6% of people have type O, Rh- blood

b. $P(\text{NOT(type O AND Rh-)}) = 1 - P(\text{type O AND Rh-}) = 1 - 0.06 = 0.94$

94% of people do not have type O, Rh- blood

Q 3.4.31

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.

Q 3.4.32

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.

1. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
2. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

S 3.4.32

a. Let $C =$ be the event that the cookie contains chocolate. Let $N =$ the event that the cookie contains nuts.
b. $P(\text{C OR N}) = P(\text{C}) + P(\text{N}) - P(\text{C AND N}) = 0.36 + 0.12 - 0.08 = 0.40$
c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(\text{C OR N}) = 1 - 0.40 = 0.60$

Q 3.4.33

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 $\text{D} =$ event that a student takes a distance learning class and $\text{E} =$ event that a student is a part time student

a. Find $P(\text{D AND E})$.
b. Find $P(\text{E}|\text{D})$.
c. Find $P(\text{D OR E})$.
d. Using an appropriate test, show whether D and E are independent.
e. Using an appropriate test, show whether D and E are mutually exclusive.

3.5: Contingency Tables

Use the information in the Table 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.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Q 3.5.1

What is the probability that a randomly selected senator has an “Other” affiliation?

S 3.5.1

0

Q 3.5.2

What is the probability that a randomly selected senator is up for reelection in November 2016?

Q 3.5.3

What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

S 3.5.3

$\frac{10}{67}$

Q 3.5.4

What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

Q 3.5.5

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?

S 3.5.5

$\frac{10}{34}$

Q 3.5.6

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?

Q 3.5.7

The events “Republican” and “Up for reelection in 2016” are _____

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

S 3.5.7

d

Q 3.5.8

The events “Other” and “Up for reelection in November 2016” are _____

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Q 3.5.9

This table 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.

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 a black or white male.
g. Out of the individuals over age 65, find the probability that a randomly selected individual was a black or white male.

S 3.5.9

a.

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

b.

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

c. $\frac{22,050}{29,760}$
d. $\frac{330}{29,760}$
e. $\frac{2,000}{29,760}$
f. $\frac{23,720}{29,760}$
g. $\frac{5,010}{6,020}$

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.

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

Q 3.5.10

Find $P(\text{hit was made by Babe Ruth})$.

a. $\frac{1518}{2873}$
b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

Q 3.5.11

Find $P(\text{hit was made by Ty Cobb}|\text{The hit was a Home Run})$.

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

S 3.5.11

b

Q 3.5.12

Table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If **B** is the event of a child having brown hair, find the probability of the complement of **B**.
g. In words, what does the complement of **B** represent?

Q 3.5.13

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.

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.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

S 3.5.13

a. $\frac{26}{106}$
b. $\frac{33}{106}$
c. $\frac{21}{106}$
d. $\left(\frac{26}{106}\right)+\left(\frac{33}{106}\right)-\left(\frac{21}{106}\right)=\left(\frac{38}{106}\right)$
e. $\frac{21}{33}$

3.6: Tree and Venn Diagrams

Exercise 3.6.8

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: $\mathrm{C}=$ a man develops cancer in his lifetime; $\mathrm{P}=$ man has at least one false positive. Construct a tree diagram of the situation.

Answer

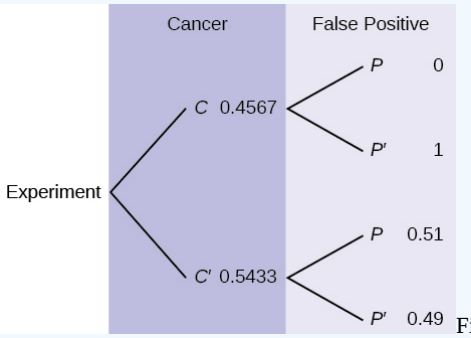


Figure 3.6.12

Bring It Together

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.

Exercise 3.6.9

Suppose that you randomly draw two cards, one at a time, **with replacement**.

Let $\mathrm{G_1}=$ first card is green
Let $\mathrm{G_2}=$ second card is green

a. Draw a tree diagram of the situation.
b. Find $P(\mathrm{G_1 AND G_2})$.

c. Find $P(\text{at least one green})$.
d. Find $P(G_2|G_1)$.
e. Are G_1 and G_2 independent events? Explain why or why not.

Answer

a.

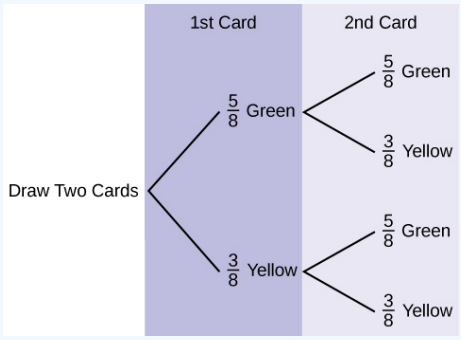


Figure 3.6.14

b. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
c. $P(\text{at least one green}) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
d. $P(G|G) = \frac{5}{8}$
e. 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.

Exercise 3.6.10

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

a. Draw a tree diagram of the situation.
b. Find $P(G_1 \text{AND } G_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(G_2|G_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

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.

Exercise 3.6.11

Complete the following.

a. Construct a table or a tree diagram of the situation.
b. Find $P(\text{driver is female})$.
c. Find $P(\text{driver is age 65 or over}|\text{driver is female})$.
d. Find $P(\text{driver is age 65 or over AND female})$.
e. In words, explain the difference between the probabilities in part c and part d.
f. Find $P(\text{driver is age 65 or over})$.
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

Answer

https://stats.libretexts.org/@go/page/48359

a.

	<20	20–64	>64	Totals
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

b. $P(\text{F}) = 0.486$
c. $P(>64|\text{F}) = 0.1361$
d. $P(>64 \text{ and } \text{F}) = P(\text{F})P(>64|\text{F}) = (0.486)(0.1361) = 0.0661$
e. $P(>64|\text{F})$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } \text{F})$ is the percentage of drivers who are female and 65 or older.
f. $P(>64) = P(>64 \text{ and } \text{F}) + P(>64 \text{ and } \text{M}) = 0.1356$
g. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } \text{F}) = 0.0661$.

Exercise 3.6.12

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.

Exercise 3.6.13

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.

a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

Answer

a.

	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

b. 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: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$.
c. Make the same assumptions as in (b) we have: (0.7708)(1,000) = 771
d. (0.1332)(1,000) = 133

Exercise 3.6.14

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.

a. Based on the given data, find $P(\mathrm{H})$ and $P(\mathrm{T})$.
b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
d. Use the tree to find the probability of obtaining at least one head.

Exercise 3.6.15

Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	____
Male	2,146	463	60	135	____
Totals	____	____	____	____	____

Suppose a person with AIDS in Santa Clara County is randomly selected.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person has a risk factor heterosexual contact})$.
c. Find $P(\text{Person is female OR has a risk factor of IV drug user})$.
d. Find $P(\text{Person is female AND has a risk factor of homosexual/bisexual})$.
e. Find $P(\text{Person is male AND has a risk factor of IV drug user})$.
f. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$.
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

Answer

The completed contingency table is as follows:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

a. $\frac{255}{2059}$
b. $\frac{196}{3059}$
c. $\frac{718}{3059}$
d. 0
e. $\frac{463}{3059}$
f. $\frac{136}{196}$

g.

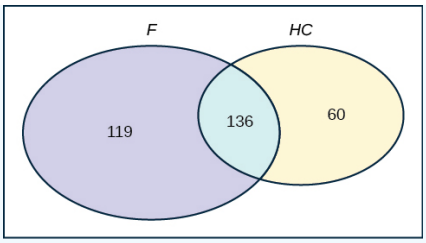


Figure 3.6.15

Exercise 3.6.16

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person obtained the disease through heterosexual contact})$.
c. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

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.

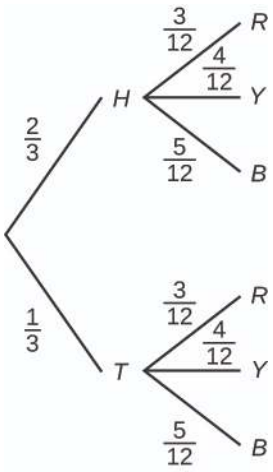


Figure 3.6.1.

Q 3.6.1

Find $P(\text{tossing a Head on the coin AND a Red bead})$

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

Q 3.6.2

Find $P(\text{Blue bead})$.

a. $\frac{15}{36}$
b. $\frac{10}{36}$

https://stats.libretexts.org/@go/page/48359

c. $\frac{10}{12}$
d. $\frac{6}{36}$

S 3.6.2

a

Q 3.6.3

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?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let $\mathbf{S}$ be the event that both cookies selected were the same flavor. Find $P(\mathbf{S})$.
e. Let $\mathbf{T}$ be the event that the cookies selected were different flavors. Find $P(\mathbf{T})$ 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.
f. Let $\mathbf{U}$ be the event that the second cookie selected is a butter cookie. Find $P(\mathbf{U})$.

3.7: Probability Topics

This page titled 4.E: Probability Topics (Exericses) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **3.E: Probability Topics (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

5: Discrete Random Variables

It is often the case that a number is naturally associated to the outcome of a random experiment: the number of boys in a three-child family, the number of defective light bulbs in a case of 100 bulbs, the length of time until the next customer arrives at the drive-through window at a bank. Such a number varies from trial to trial of the corresponding experiment, and does so in a way that cannot be predicted with certainty; hence, it is called a random variable. In this chapter and the next we study such variables.

5.1: Random Variables

5.1.1: Probability Distributions for Discrete Random Variables

5.2: The Binomial Distribution

5.E: Discrete Random Variables (Exercises)

This page titled 5: Discrete Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

5.1: Random Variables

Learning Objectives

- To learn the concept of a random variable.
- To learn the distinction between discrete and continuous random variables.

Definition: random variable

A random variable is a numerical quantity that is generated by a random experiment.

We will denote random variables by capital letters, such as X or Z, and the actual values that they can take by lowercase letters, such as x and z.

Table 5.1.1 gives four examples of random variables. In the second example, the three dots indicates that every counting number is a possible value for X. Although it is highly unlikely, for example, that it would take 50 tosses of the coin to observe heads for the first time, nevertheless it is conceivable, hence the number 50 is a possible value. The set of possible values is infinite, but is still at least countable, in the sense that all possible values can be listed one after another. In the last two examples, by way of contrast, the possible values cannot be individually listed, but take up a whole interval of numbers. In the fourth example, since the light bulb could conceivably continue to shine indefinitely, there is no natural greatest value for its lifetime, so we simply place the symbol ∞ for infinity as the right endpoint of the interval of possible values.

Table 5.1.1: Four Random Variables

Experiment	Number X	Possible Values of X
Roll two fair dice	Sum of the number of dots on the top faces	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Flip a fair coin repeatedly	Number of tosses until the coin lands heads	1, 2, 3,4, …
Measure the voltage at an electrical outlet	Voltage measured	$118 \le x \le 122$
Operate a light bulb until it burns out	Time until the bulb burns out	$0 \le x < \infty$

Definition: discrete random variable

A random variable is called discrete if it has either a finite or a countable number of possible values. A random variable is called continuous if its possible values contain a whole interval of numbers.

The examples in the table are typical in that discrete random variables typically arise from a counting process, whereas continuous random variables typically arise from a measurement.

Key Takeaway

- A random variable is a number generated by a random experiment.
- A random variable is called discrete if its possible values form a finite or countable set.
- A random variable is called continuous if its possible values contain a whole interval of numbers.

This page titled 5.1: Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **4.1: Random Variables** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

5.1.1: Probability Distributions for Discrete Random Variables

Learning Objectives

- To learn the concept of the probability distribution of a discrete random variable.
- To learn the concepts of the mean, variance, and standard deviation of a discrete random variable, and how to compute them.

Associated to each possible value x of a discrete random variable X is the probability $P(x)$ that X will take the value x in one trial of the experiment.

Definition: probability distribution

The probability distribution of a discrete random variable X is a list of each possible value of X together with the probability that X takes that value in one trial of the experiment.

The probabilities in the probability distribution of a random variable X must satisfy the following two conditions:

- Each probability $P(x)$ must be between 0 and 1:

$$0 \le P(x) \le 1.$$

- The sum of all the possible probabilities is 1:

$$\sum P(x) = 1.$$

Example 5.1.1.1: two Fair Coins

A fair coin is tossed twice. Let X be the number of heads that are observed.

a. Construct the probability distribution of X.
b. Find the probability that at least one head is observed.

Solution

a. The possible values that X can take are 0, 1, and 2. Each of these numbers corresponds to an event in the sample space $S = \{hh, ht, th, tt\}$ of equally likely outcomes for this experiment:

$$X = 0 \text{ to } \{tt\},\ X = 1 \text{ to } \{ht, th\},\ \text{and } X = 2 \text{ to } hh.$$

The probability of each of these events, hence of the corresponding value of X, can be found simply by counting, to give

x	0	1	2
$P(x)$	0.25	0.50	0.25

This table is the probability distribution of X.

b. "At least one head" is the event $X \ge 1$, which is the union of the mutually exclusive events $X = 1$ and $X = 2$. Thus

$$P(X \ge 1) = P(1) + P(2) = 0.50 + 0.25$$
$$= 0.75$$

A histogram that graphically illustrates the probability distribution is given in Figure 5.1.1.1.

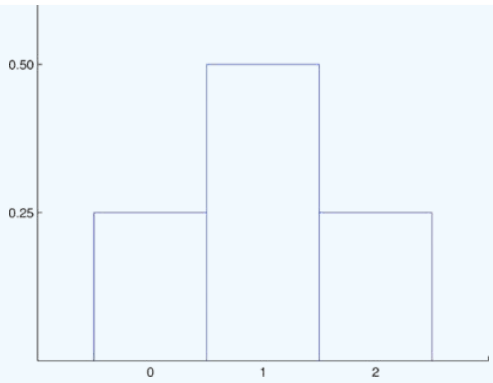


Figure 5.1.1.1: *Probability Distribution for Tossing a Fair Coin Twice*

Example 5.1.1.2: Two Fair Dice

A pair of fair dice is rolled. Let X denote the sum of the number of dots on the top faces.

a. Construct the probability distribution of X for a paid of fair dice.
b. Find $P(X \geq 9)$.
c. Find the probability that X takes an even value.

Solution

The sample space of equally likely outcomes is

$$\begin{array}{cccccc} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array}$$

where the first digit is die 1 and the second number is die 2.

a. The possible values for X are the numbers 2 through 12. $X = 2$ is the event $\{11\}$, so $P(2) = 1/36$. $X = 3$ is the event $\{12, 21\}$, so $P(3) = 2/36$. Continuing this way we obtain the following table

x	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This table is the probability distribution of X.

b. The event $X \geq 9$ is the union of the mutually exclusive events $X = 9$, $X = 10$, $X = 11$, and $X = 12$. Thus

$$\begin{aligned} P(X \geq 9) &= P(9) + P(10) + P(11) + P(12) \\ &= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} \\ &= \frac{10}{36} \\ &= 0.2\overline{7} \end{aligned}$$

c. Before we immediately jump to the conclusion that the probability that X takes an even value must be 0.5, note that X takes six different even values but only five different odd values. We compute

$$
\begin{aligned}
P(X \text{ is even}) &= P(2) + P(4) + P(6) + P(8) + P(10) + P(12) \\
&= \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} \\
&= \frac{18}{36} \\
&= 0.5
\end{aligned}
$$

A histogram that graphically illustrates the probability distribution is given in Figure 5.1.1.2.

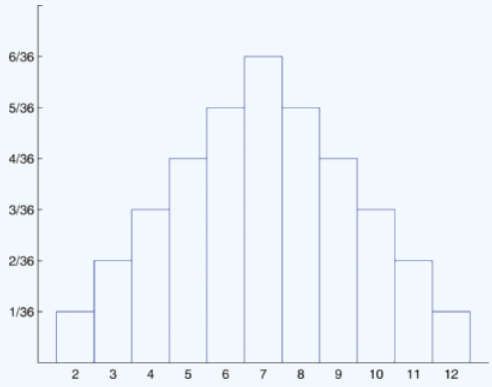


Figure 5.1.1.2: *Probability Distribution for Tossing Two Fair Dice*

The Mean and Standard Deviation of a Discrete Random Variable

Definition: mean

The *mean* (also called the "expectation value" or "expected value") of a discrete random variable X is the number

$$\mu = E(X) = \sum xP(x) \tag{5.1.1.1}$$

The mean of a random variable may be interpreted as the average of the values assumed by the random variable in repeated trials of the experiment.

Example 5.1.1.3

Find the mean of the discrete random variable X whose probability distribution is

x	-2	1	2	3.5
$P(x)$	0.21	0.34	0.24	0.21

Solution

Using the definition of mean (Equation 5.1.1.1) gives

$$
\begin{aligned}
\mu &= \sum xP(x) \\
&= (-2)(0.21) + (1)(0.34) + (2)(0.24) + (3.5)(0.21) \\
&= 1.135
\end{aligned}
$$

Example 5.1.1.4

A service organization in a large town organizes a raffle each month. One thousand raffle tickets are sold for $1 each. Each has an equal chance of winning. First prize is $300, second prize is $200, and third prize is $100. Let X denote the net gain from the purchase of one ticket.

a. Construct the probability distribution of X.
b. Find the probability of winning any money in the purchase of one ticket.
c. Find the expected value of X, and interpret its meaning.

Solution

a. If a ticket is selected as the first prize winner, the net gain to the purchaser is the $300 prize less the $1 that was paid for the ticket, hence $X = 300 - 11 = 299$. There is one such ticket, so $P(299) = 0.001$. Applying the same "income minus outgo" principle to the second and third prize winners and to the 997 losing tickets yields the probability distribution:

x	299	199	99	-1
$P(x)$	0.001	0.001	0.001	0.997

b. Let W denote the event that a ticket is selected to win one of the prizes. Using the table

$$\begin{aligned} P(W) &= P(299) + P(199) + P(99) = 0.001 + 0.001 + 0.001 \\ &= 0.003 \end{aligned}$$

c. Using the definition of expected value (Equation 5.1.1.1),

$$\begin{aligned} E(X) &= (299) \cdot (0.001) + (199) \cdot (0.001) + (99) \cdot (0.001) + (-1) \cdot (0.997) \\ &= -0.4 \end{aligned}$$

The negative value means that one loses money on the average. In particular, if someone were to buy tickets repeatedly, then although he would win now and then, on average he would lose 40 cents per ticket purchased.

The concept of expected value is also basic to the insurance industry, as the following simplified example illustrates.

Example 5.1.1.5

A life insurance company will sell a $200,000 one-year term life insurance policy to an individual in a particular risk group for a premium of $195. Find the expected value to the company of a single policy if a person in this risk group has a 99.97% chance of surviving one year.

Solution

Let X denote the net gain to the company from the sale of one such policy. There are two possibilities: the insured person lives the whole year or the insured person dies before the year is up. Applying the "income minus outgo" principle, in the former case the value of X is $195 - 0$; in the latter case it is $195 - 200,000 = -199,805$. Since the probability in the first case is 0.9997 and in the second case is $1 - 0.9997 = 0.0003$, the probability distribution for X is:

x	195	$-199,805$
$P(x)$	0.9997	0.0003

Therefore

$$\begin{aligned} E(X) &= \sum xP(x) \\ &= (195) \cdot (0.9997) + (-199,805) \cdot (0.0003) \\ &= 135 \end{aligned}$$

Occasionally (in fact, 3 times in 10,000) the company loses a large amount of money on a policy, but typically it gains $195, which by our computation of $E(X)$ works out to a net gain of $135 per policy sold, on average.

Definition: variance

The *variance* (σ^2) of a discrete random variable X is the number

$$\sigma^2 = \sum (x - \mu)^2 P(x) \tag{5.1.1.2}$$

which by algebra is equivalent to the formula

$$\sigma^2 = \left[\sum x^2 P(x)\right] - \mu^2 \tag{5.1.1.3}$$

Definition: standard deviation

The standard deviation, σ, *of a discrete random variable* X *is the square root of its variance, hence is given by the formulas*

$$\sigma = \sqrt{\sum (x-\mu)^2 P(x)} = \sqrt{\left[\sum x^2 P(x)\right] - \mu^2} \tag{5.1.1.4}$$

The variance and standard deviation of a discrete random variable X may be interpreted as measures of the variability of the values assumed by the random variable in repeated trials of the experiment. The units on the standard deviation match those of X.

Example 5.1.1.6

A discrete random variable X has the following probability distribution:

x	-1	0	1	4
$P(x)$	0.2	0.5	a	0.1

(5.1.1.5)

A histogram that graphically illustrates the probability distribution is given in Figure 5.1.1.3.

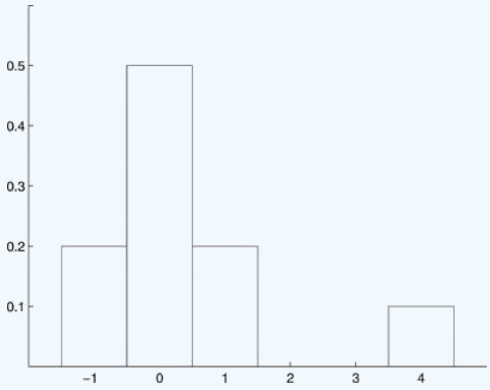


Figure 5.1.1.3: Probability Distribution of a Discrete Random Variable

Compute each of the following quantities.

a. a.
b. $P(0)$.
c. $P(X > 0)$.
d. $P(X \geq 0)$.
e. $P(X \leq -2)$.
f. The mean μ of X.
g. The variance σ^2 of X.
h. The standard deviation σ of X.

Solution

a. Since all probabilities must add up to 1,

$$a = 1 - (0.2 + 0.5 + 0.1) = 0.2$$

b. Directly from the table, P(0)=0.5

$$P(0) = 0.5$$

c. From Table 5.1.1.5,

$$P(X > 0) = P(1) + P(4) = 0.2 + 0.1 = 0.3$$

d. From Table 5.1.1.5,

$$P(X \geq 0) = P(0) + P(1) + P(4) = 0.5 + 0.2 + 0.1 = 0.8$$

e. Since none of the numbers listed as possible values for X is less than or equal to -2, the event $X \leq -2$ is impossible, so

$$P(X \leq -2) = 0$$

f. Using the formula in the definition of μ (Equation 5.1.1.1)

$$\begin{aligned}\mu &= \sum xP(x) \\ &= (-1) \cdot (0.2) + (0) \cdot (0.5) + (1) \cdot (0.2) + (4) \cdot (0.1) \\ &= 0.4\end{aligned}$$

g. Using the formula in the definition of σ^2 (Equation 5.1.1.2) and the value of μ that was just computed,

$$\begin{aligned}\sigma^2 &= \sum (x-\mu)^2 P(x) \\ &= (-1-0.4)^2 \cdot (0.2) + (0-0.4)^2 \cdot (0.5) + (1-0.4)^2 \cdot (0.2) + (4-0.4)^2 \cdot (0.1) \\ &= 1.84\end{aligned}$$

h. Using the result of part (g), $\sigma = \sqrt{1.84} = 1.3565$

Summary

- The probability distribution of a discrete random variable X is a listing of each possible value x taken by X along with the probability $P(x)$ that X takes that value in one trial of the experiment.
- The mean μ of a discrete random variable X is a number that indicates the average value of X over numerous trials of the experiment. It is computed using the formula $\mu = \sum xP(x)$.
- The variance σ^2 and standard deviation σ of a discrete random variable X are numbers that indicate the variability of X over numerous trials of the experiment. They may be computed using the formula $\sigma^2 = [\sum x^2 P(x)] - \mu^2$.

This page titled 5.1.1: Probability Distributions for Discrete Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **4.2: Probability Distributions for Discrete Random Variables** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

5.2: The Binomial Distribution

Learning Objectives

- To learn the concept of a binomial random variable.
- To learn how to recognize a random variable as being a binomial random variable.

The experiment of tossing a fair coin three times and the experiment of observing the genders according to birth order of the children in a randomly selected three-child family are completely different, but the random variables that count the number of heads in the coin toss and the number of boys in the family (assuming the two genders are equally likely) are the same random variable, the one with probability distribution

x	0	1	2	3
$P(x)$	0.125	0.375	0.375	0.125

A histogram that graphically illustrates this probability distribution is given in Figure 5.2.1. What is common to the two experiments is that we perform three identical and independent trials of the same action, each trial has only two outcomes (heads or tails, boy or girl), and the probability of success is the same number, 0.5, on every trial. The random variable that is generated is called the binomial random variable with parameters $n = 3$ and $p = 0.5$. This is just one case of a general situation.

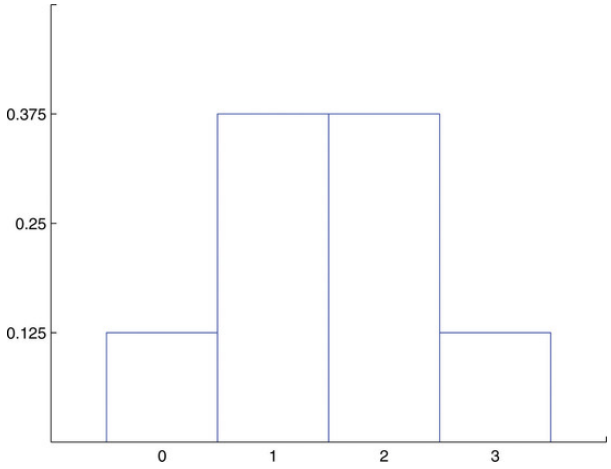


Figure 5.2.1: Probability Distribution for Three Coins and Three Children

Definition: binomial distribution

Suppose a random experiment has the following characteristics.

- There are n identical and independent trials of a common procedure.
- There are exactly two possible outcomes for each trial, one termed "success" and the other "failure."
- The probability of success on any one trial is the same number p.

Then the discrete random variable X that counts the number of successes in the n trials is the binomial random variable with parameters n and p. We also say that X has a binomial distribution with parameters n and p.

The following four examples illustrate the definition. Note how in every case "success" is the outcome that is counted, not the outcome that we prefer or think is better in some sense.

1. A random sample of 125 students is selected from a large college in which the proportion of students who are females is 57%. Suppose X denotes the number of female students in the sample. In this situation there are $n = 125$ identical and independent trials of a common procedure, selecting a student at random; there are exactly two possible outcomes for each trial, "success" (what we are counting, that the student be female) and "failure;" and finally the probability of success on any one trial is the same number $p = 0.57$. X is a binomial random variable with parameters $n = 125$ and $p = 0.57$.

2. A multiple-choice test has 15 questions, each of which has five choices. An unprepared student taking the test answers each of the questions completely randomly by choosing an arbitrary answer from the five provided. Suppose X denotes the number of answers that the student gets right. X is a binomial random variable with parameters $n = 15$ and $p = 1/5 = 0.20$.
3. In a survey of 1,000 registered voters each voter is asked if he intends to vote for a candidate Titania Queen in the upcoming election. Suppose X denotes the number of voters in the survey who intend to vote for Titania Queen. X is a binomial random variable with $n = 1000$ and p equal to the true proportion of voters (surveyed or not) who intend to vote for Titania Queen.
4. An experimental medication was given to 30 patients with a certain medical condition. Suppose X denotes the number of patients who develop severe side effects. X is a binomial random variable with $n = 30$ and p equal to the true probability that a patient with the underlying condition will experience severe side effects if given that medication.

Probability Formula for a Binomial Random Variable

Often the most difficult aspect of working a problem that involves the binomial random variable is recognizing that the random variable in question has a binomial distribution. Once that is known, probabilities can be computed using the following formula.

If X is a binomial random variable with parameters n and p, then

$$P(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

where $q = 1 - p$ and where for any counting number m, $m!$ (read "m factorial") is defined by

$$0! = 1, 1! = 1, 2! = 1 \cdot 2, 3! = 1 \cdot 2 \cdot 3$$

and in general

$$m! = 1 \cdot 2 \cdots (m-1) \cdot m$$

Example 5.2.1

Seventeen percent of victims of financial fraud know the perpetrator of the fraud personally.

a. Use the formula to construct the probability distribution for the number X of people in a random sample of five victims of financial fraud who knew the perpetrator personally.
b. A investigator examines five cases of financial fraud every day. Find the most frequent number of cases each day in which the victim knew the perpetrator.
c. A investigator examines five cases of financial fraud every day. Find the average number of cases per day in which the victim knew the perpetrator.

Solution

The random variable X is binomial with parameters $n = 5$ and $p = 0.17$; $q = 1 - p = 0.83$. The possible values of X are $0, 1, 2, 3, 4,$ and 5.

$$\begin{aligned} P(0) &= \frac{5!}{0!5!}(0.17)^0(0.83)^5 \\ &= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)} 1 \cdot (0.3939040643) \\ &= 0.3939040643 \approx 0.3939 \end{aligned}$$

$$\begin{aligned} P(1) &= \frac{5!}{1!4!}(0.17)^1(0.83)^4 \\ &= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4)}(0.17) \cdot (0.47458321) \\ &= 5 \cdot (0.17) \cdot (0.47458321) \\ &= 0.4033957285 \approx 0.4034 \end{aligned}$$

$$
\begin{aligned}
P(2) &= \frac{5!}{2!3!}(0.17)^2(0.83)^3 \\
&= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1 \cdot 2)(1 \cdot 2 \cdot 3)}(0.0289) \cdot (0.571787) \\
&= 10 \cdot (0.0289) \cdot (0.571787) \\
&= 0.165246443 \approx 0.1652
\end{aligned}
$$

The remaining three probabilities are computed similarly, to give the probability distribution

x	0	1	2	3	4	5
$P(x)$	0.3939	0.4034	0.1652	0.0338	0.0035	0.0001

The probabilities do not add up to exactly 1 because of rounding.

This probability distribution is represented by the histogram in Figure 5.2.2, which graphically illustrates just how improbable the events $X = 4$ and $X = 5$ are. The corresponding bar in the histogram above the number 4 is barely visible, if visible at all, and the bar above 5 is far too short to be visible.

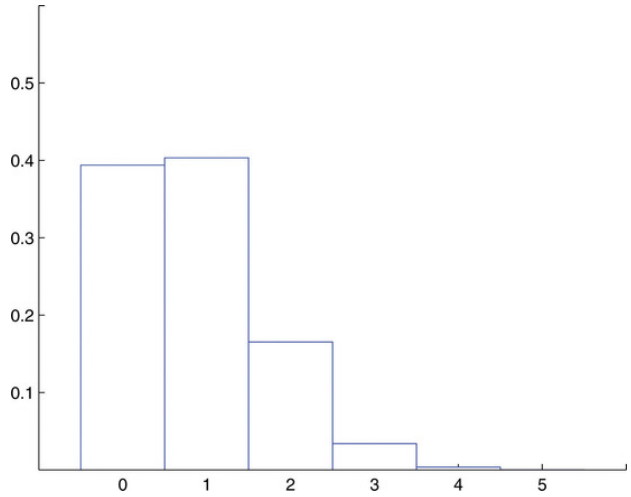


Figure 5.2.2: Probability Distribution of the Binomial Random Variable in Example 5.2.1

The value of X that is most likely is $X = 1$, so the most frequent number of cases seen each day in which the victim knew the perpetrator is one.

The average number of cases per day in which the victim knew the perpetrator is the mean of X, which is

$$
\begin{aligned}
\mu &= \sum xP(x) && (5.2.1) \\
&= 0 \cdot 0.3939 + 1 \cdot 0.4034 + 2 \cdot 0.1652 + 3 \cdot 0.0338 + 4 \cdot 0.0035 + 5 \cdot 0.0001 && (5.2.2) \\
&= 0.8497 && (5.2.3)
\end{aligned}
$$

Special Formulas for the Mean and Standard Deviation of a Binomial Random Variable

Since a binomial random variable is a discrete random variable, the formulas for its mean, variance, and standard deviation given in the previous section apply to it, as we just saw in Example 5.2.2 in the case of the mean. However, for the binomial random variable there are much simpler formulas.

If X is a binomial random variable with parameters n and p, then

$$\mu = np$$

$$\sigma^2 = npq$$

$$\sigma = \sqrt{npq}$$

where $q = 1 - p$.

Example 5.2.2

Find the mean and standard deviation of the random variable X of Example 5.2.1.

Solution

The random variable X is binomial with parameters $n = 5$ and $p = 0.17$, and $q = 1 - p = 0.83$. Thus its mean and standard deviation are

$$\mu = np = (5) \cdot (0.17) = 0.85 \text{ (exactly)}$$

and

$$\sigma = \sqrt{npq} = \sqrt{(5) \cdot (0.17) \cdot (0.83)} = \sqrt{0.7055} \approx 0.8399$$

The Cumulative Probability Distribution of a Binomial Random Variable

In order to allow a broader range of more realistic problems contains probability tables for binomial random variables for various choices of the parameters n and p. These tables are not the probability distributions that we have seen so far, but are **cumulative probability distributions**. In the place of the probability $P(x)$ the table contains the probability

$$P(X \leq x) = P(0) + P(1) + ... + P(x)$$

This is illustrated in Figure 5.2.3. The probability entered in the table corresponds to the area of the shaded region. The reason for providing a cumulative table is that in practical problems that involve a binomial random variable typically the probability that is sought is of the form $P(X \leq x)$ or $P(X \geq x)$. The cumulative table is much easier to use for computing $P(X \leq x)$ since all the individual probabilities have already been computed and added. The one table suffices for both $P(X \leq x)$ or $P(X \geq x)$ and can be used to readily obtain probabilities of the form $P(x)$, too, because of the following formulas. The first is just the Probability Rule for Complements.

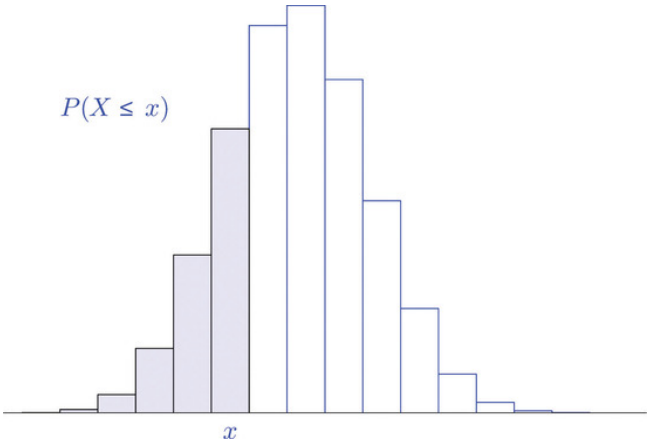


Figure 5.2.3: Cumulative Probabilities

If X is a discrete random variable, then

$$P(X \geq x) = 1 - P(X \leq x - 1)$$

and

$$P(x) = P(X \leq x) - P(X \leq x - 1)$$

Example 5.2.3

A student takes a ten-question true/false exam.

a. Find the probability that the student gets exactly six of the questions right simply by guessing the answer on every question.
b. Find the probability that the student will obtain a passing grade of 60% or greater simply by guessing.

Solution

Let X denote the number of questions that the student guesses correctly. Then X is a binomial random variable with parameters $n = 10$ and $p = 0.50$.

a. The probability sought is $P(6)$. The formula gives

$$P(6) = 10!(6!)(4!)(.5)6.54 = 0.205078125$$

Using the table,

$$P(6) = P(X \leq 6) - P(X \leq 5) = 0.8281 - 0.6230 = 0.2051$$

b. The student must guess correctly on at least 60% of the questions, which is $(0.60) \cdot (10) = 6$ questions. The probability sought is not $P(6)$ (an easy mistake to make), but

$$P(X \geq 6) = P(6) + P(7) + P(8) + P(9) + P(10)$$

Instead of computing each of these five numbers using the formula and adding them we can use the table to obtain

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.6230 = 0.3770$$

which is much less work and of sufficient accuracy for the situation at hand.

Example 5.2.4

An appliance repairman services five washing machines on site each day. One-third of the service calls require installation of a particular part.

a. The repairman has only one such part on his truck today. Find the probability that the one part will be enough today, that is, that at most one washing machine he services will require installation of this particular part.
b. Find the minimum number of such parts he should take with him each day in order that the probability that he have enough for the day's service calls is at least 95%.

Solution

Let X denote the number of service calls today on which the part is required. Then X is a binomial random variable with parameters $n = 5$ and $p = 1/3 = 0.\bar{3}$

a. Note that the probability in question is not $P(1)$, but rather $P(X \leq 1)$. Using the cumulative distribution table,

$$P(X \leq 1) = 0.4609$$

b. The answer is the smallest number x such that the table entry $P(X \leq x)$ is at least 0.9500. Since $P(X \leq 2) = 0.7901$ is less than 0.95, two parts are not enough. Since $P(X \leq 3) = 0.9547$ is as large as 0.95, three parts will suffice at least 95% of the time. Thus the minimum needed is three.

Summary

- The discrete random variable X that counts the number of successes in n identical, independent trials of a procedure that always results in either of two outcomes, “success” or “failure,” and in which the probability of success on each trial is the same number p, is called the binomial random variable with parameters n and p.
- There is a formula for the probability that the binomial random variable with parameters n and p will take a particular value x.
- There are special formulas for the mean, variance, and standard deviation of the binomial random variable with parameters n and p that are much simpler than the general formulas that apply to all discrete random variables.
- Cumulative probability distribution tables, when available, facilitate computation of probabilities encountered in typical practical situations.

This page titled 5.2: The Binomial Distribution is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **4.3: The Binomial Distribution** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

5.E: Discrete Random Variables (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

4.1: Random Variables

Basic

1. Classify each random variable as either discrete or continuous.
   a. The number of arrivals at an emergency room between midnight and $6:00\ a.m.$
   b. The weight of a box of cereal labeled "18 ounces."
   c. The duration of the next outgoing telephone call from a business office.
   d. The number of kernels of popcorn in a 1-pound container.
   e. The number of applicants for a job.
2. Classify each random variable as either discrete or continuous.
   a. The time between customers entering a checkout lane at a retail store.
   b. The weight of refuse on a truck arriving at a landfill.
   c. The number of passengers in a passenger vehicle on a highway at rush hour.
   d. The number of clerical errors on a medical chart.
   e. The number of accident-free days in one month at a factory.
3. Classify each random variable as either discrete or continuous.
   a. The number of boys in a randomly selected three-child family.
   b. The temperature of a cup of coffee served at a restaurant.
   c. The number of no-shows for every 100 reservations made with a commercial airline.
   d. The number of vehicles owned by a randomly selected household.
   e. The average amount spent on electricity each July by a randomly selected household in a certain state.
4. Classify each random variable as either discrete or continuous.
   a. The number of patrons arriving at a restaurant between $5:00\ p.m.$ and $6:00\ p.m.$
   b. The number of new cases of influenza in a particular county in a coming month.
   c. The air pressure of a tire on an automobile.
   d. The amount of rain recorded at an airport one day.
   e. The number of students who actually register for classes at a university next semester.
5. Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)
   a. The number of heads in two tosses of a coin.
   b. The average weight of newborn babies born in a particular county one month.
   c. The amount of liquid in a 12-ounce can of soft drink.
   d. The number of games in the next World Series (best of up to seven games).
   e. The number of coins that match when three coins are tossed at once.
6. Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)
   a. The number of hearts in a five-card hand drawn from a deck of 52 cards that contains 13 hearts in all.
   b. The number of pitches made by a starting pitcher in a major league baseball game.
   c. The number of breakdowns of city buses in a large city in one week.
   d. The distance a rental car rented on a daily rate is driven each day.
   e. The amount of rainfall at an airport next month.

Answers

1. a. discrete
   b. continuous

c. continuous
d. discrete
e. discrete

2.

3. a. discrete
   b. continuous
   c. discrete
   d. discrete
   e. continuous

4.

5. a. $\{0.1.2\}$
   b. an interval (a, b) (answers vary)
   c. an interval (a, b) (answers vary)
   d. $\{4, 5, 6, 7\}$
   e. $\{2, 3\}$

4.2: Probability Distributioins for Discrete Random Variables

Basic

1. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.
   $$\begin{array}{c|cccc} x & -2 & 0 & 2 & 4 \\ \hline P(x) & 0.3 & 0.5 & 0.2 & 0.1 \end{array} \tag{5.E.1}$$

   b.
   $$\begin{array}{c|ccc} x & 0.5 & 0.25 & 0.25 \\ \hline P(x) & -0.4 & 0.6 & 0.8 \end{array} \tag{5.E.2}$$

   c.
   $$\begin{array}{c|ccccc} x & 1.1 & 2.5 & 4.1 & 4.6 & 5.3 \\ \hline P(x) & 0.16 & 0.14 & 0.11 & 0.27 & 0.22 \end{array} \tag{5.E.3}$$

2. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.
   $$\begin{array}{c|ccccc} x & 0 & 1 & 2 & 3 & 4 \\ \hline P(x) & -0.25 & 0.50 & 0.35 & 0.10 & 0.30 \end{array} \tag{5.E.4}$$

   b.
   $$\begin{array}{c|ccc} x & 1 & 2 & 3 \\ \hline P(x) & 0.325 & 0.406 & 0.164 \end{array} \tag{5.E.5}$$

   c.
   $$\begin{array}{c|ccccc} x & 25 & 26 & 27 & 28 & 29 \\ \hline P(x) & 0.13 & 0.27 & 0.28 & 0.18 & 0.14 \end{array} \tag{5.E.6}$$

3. A discrete random variable X has the following probability distribution:

   $$\begin{array}{c|ccccc} x & 77 & 78 & 79 & 80 & 81 \\ \hline P(x) & 0.15 & 0.15 & 0.20 & 0.40 & 0.10 \end{array} \tag{5.E.7}$$

   Compute each of the following quantities.

   a. $P(80)$.
   b. $P(X > 80)$.
   c. $P(X \leq 80)$.
   d. The mean μ of X.
   e. The variance σ^2 of X.
   f. The standard deviation σ of X.

4. A discrete random variable X has the following probability distribution:

x	13	18	20	24	27
$P(x)$	0.22	0.25	0.20	0.17	0.16

(5.E.8)

Compute each of the following quantities.

a. $P(18)$.
b. $P(X > 18)$.
c. $P(X \leq 18)$.
d. The mean μ of X.
e. The variance σ^2 of X.
f. The standard deviation σ of X.

5. If each die in a pair is "loaded" so that one comes up half as often as it should, six comes up half again as often as it should, and the probabilities of the other faces are unaltered, then the probability distribution for the sum X of the number of dots on the top faces when the two are rolled is

x	2	3	4	5	6	7
$P(x)$	$\frac{1}{144}$	$\frac{4}{144}$	$\frac{8}{144}$	$\frac{12}{144}$	$\frac{16}{144}$	$\frac{22}{144}$

(5.E.9)

x	8	9	10	11	12
$P(x)$	$\frac{24}{144}$	$\frac{20}{144}$	$\frac{16}{144}$	$\frac{12}{144}$	$\frac{9}{144}$

(5.E.10)

Compute each of the following.

a. $P(5 \leq X \leq 9)$.
b. $P(X \geq 7)$.
c. The mean μ of X. (For fair dice this number is 7).
d. The standard deviation σ of X. (For fair dice this number is about 2.415).

Applications

6. Borachio works in an automotive tire factory. The number X of sound but blemished tires that he produces on a random day has the probability distribution

x	2	3	4	5
$P(x)$	0.48	0.36	0.12	0.04

(5.E.11)

a. Find the probability that Borachio will produce more than three blemished tires tomorrow.
b. Find the probability that Borachio will produce at most two blemished tires tomorrow.
c. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

7. In a hamster breeder's experience the number X of live pups in a litter of a female not over twelve months in age who has not borne a litter in the past six weeks has the probability distribution

x	3	4	5	6	7	8	9
$P(x)$	0.04	0.10	0.26	0.31	0.22	0.05	0.02

(5.E.12)

a. Find the probability that the next litter will produce five to seven live pups.
b. Find the probability that the next litter will produce at least six live pups.
c. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

8. The number X of days in the summer months that a construction crew cannot work because of the weather has the probability distribution

x	6	7	8	9	10
$P(x)$	0.03	0.08	0.15	0.20	0.19

(5.E.13)

x	11	12	13	14
$P(x)$	0.16	0.10	0.07	0.02

(5.E.14)

a. Find the probability that no more than ten days will be lost next summer.

b. Find the probability that from 8 to 12 days will be lost next summer.
c. Find the probability that no days at all will be lost next summer.
d. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

9. Let X denote the number of boys in a randomly selected three-child family. Assuming that boys and girls are equally likely, construct the probability distribution of X.
10. Let X denote the number of times a fair coin lands heads in three tosses. Construct the probability distribution of X.
11. Five thousand lottery tickets are sold for $\$1$ each. One ticket will win $\$1,000$, two tickets will win $\$500$ each, and ten tickets will win $\$100$ each. Let X denote the net gain from the purchase of a randomly selected ticket.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X. Interpret its meaning.
    c. Compute the standard deviation σ of X.
12. Seven thousand lottery tickets are sold for $\$5$ each. One ticket will win $\$2,000$, two tickets will win $\$750$ each, and five tickets will win $\$100$ each. Let X denote the net gain from the purchase of a randomly selected ticket.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X. Interpret its meaning.
    c. Compute the standard deviation σ of X.
13. An insurance company will sell a $\$90,000$ one-year term life insurance policy to an individual in a particular risk group for a premium of $\$478$. Find the expected value to the company of a single policy if a person in this risk group has a 99.62% chance of surviving one year.
14. An insurance company will sell a $\$10,000$ one-year term life insurance policy to an individual in a particular risk group for a premium of $\$368$. Find the expected value to the company of a single policy if a person in this risk group has a 97.25% chance of surviving one year.
15. An insurance company estimates that the probability that an individual in a particular risk group will survive one year is 0.9825. Such a person wishes to buy a $\$150,000$ one-year term life insurance policy. Let C denote how much the insurance company charges such a person for such a policy.
    a. Construct the probability distribution of X. (Two entries in the table will contain C).
    b. Compute the expected value $E(X)$ of X.
    c. Determine the value C must have in order for the company to break even on all such policies (that is, to average a net gain of zero per policy on such policies).
    d. Determine the value C must have in order for the company to average a net gain of $\$250$ per policy on all such policies.
16. An insurance company estimates that the probability that an individual in a particular risk group will survive one year is 0.99. Such a person wishes to buy a $\$75,000$ one-year term life insurance policy. Let C denote how much the insurance company charges such a person for such a policy.
    a. Construct the probability distribution of X. (Two entries in the table will contain C).
    b. Compute the expected value $E(X)$ of X.
    c. Determine the value C must have in order for the company to break even on all such policies (that is, to average a net gain of zero per policy on such policies).
    d. Determine the value C must have in order for the company to average a net gain of $\$150$ per policy on all such policies.
17. A roulette wheel has 38 slots. Thirty-six slots are numbered from 1 to 36; half of them are red and half are black. The remaining two slots are numbered 0 and 00 and are green. In a $\$1$ bet on red, the bettor pays $\$1$ to play. If the ball lands in a red slot, he receives back the dollar he bet plus an additional dollar. If the ball does not land on red he loses his dollar. Let X denote the net gain to the bettor on one play of the game.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X, and interpret its meaning in the context of the problem.
    c. Compute the standard deviation of X.
18. A roulette wheel has 38 slots. Thirty-six slots are numbered from 1 to 36; the remaining two slots are numbered 0 and 00. Suppose the "number" 00 is considered not to be even, but the number 0 is still even. In a $\$1$ bet on even, the bettor pays $\$1$ to

play. If the ball lands in an even numbered slot, he receives back the dollar he bet plus an additional dollar. If the ball does not land on an even numbered slot, he loses his dollar. Let X denote the net gain to the bettor on one play of the game.

a. Construct the probability distribution of X.
b. Compute the expected value $E(X)$ of X, and explain why this game is not offered in a casino (where 0 is not considered even).
c. Compute the standard deviation of X.

19. The time, to the nearest whole minute, that a city bus takes to go from one end of its route to the other has the probability distribution shown. As sometimes happens with probabilities computed as empirical relative frequencies, probabilities in the table add up only to a value other than 1.00 because of round-off error.

x	42	43	44	45	46	47
$P(x)$	0.10	0.23	0.34	0.25	0.05	0.02

(5.E.15)

a. Find the average time the bus takes to drive the length of its route.
b. Find the standard deviation of the length of time the bus takes to drive the length of its route.

20. Tybalt receives in the mail an offer to enter a national sweepstakes. The prizes and chances of winning are listed in the offer as: $\$5$ million, one chance in 65 million; $\$150,000$, one chance in 6.5 million; $\$5,000$, one chance in $650,000$; and $\$1,000$, one chance in $65,000$. If it costs Tybalt 44 cents to mail his entry, what is the expected value of the sweepstakes to him?

Additional Exercises

21. The number X of nails in a randomly selected 1-pound box has the probability distribution shown. Find the average number of nails per pound.

x	100	101	102
$P(x)$	0.01	0.96	0.03

(5.E.16)

22. Three fair dice are rolled at once. Let X denote the number of dice that land with the same number of dots on top as at least one other die. The probability distribution for X is

x	0	u	3
$P(x)$	p	$\frac{15}{36}$	$\frac{1}{36}$

(5.E.17)

a. Find the missing value u of X.
b. Find the missing probability p.
c. Compute the mean of X.
d. Compute the standard deviation of X.

23. Two fair dice are rolled at once. Let X denote the difference in the number of dots that appear on the top faces of the two dice. Thus for example if a one and a five are rolled, $X = 4$, and if two sixes are rolled, $X = 0$.

a. Construct the probability distribution for X.
b. Compute the mean μ of X.
c. Compute the standard deviation σ of X.

24. A fair coin is tossed repeatedly until either it lands heads or a total of five tosses have been made, whichever comes first. Let X denote the number of tosses made.

a. Construct the probability distribution for X.
b. Compute the mean μ of X.
c. Compute the standard deviation σ of X.

25. A manufacturer receives a certain component from a supplier in shipments of 100 units. Two units in each shipment are selected at random and tested. If either one of the units is defective the shipment is rejected. Suppose a shipment has 5 defective units.

a. Construct the probability distribution for the number X of defective units in such a sample. (A tree diagram is helpful).
b. Find the probability that such a shipment will be accepted.

26. Shylock enters a local branch bank at $4:30\ p.m.$ every payday, at which time there are always two tellers on duty. The number X of customers in the bank who are either at a teller window or are waiting in a single line for the next available teller has the following probability distribution.

x	0	1	2	3
$P(x)$	0.135	0.192	0.284	0.230

(5.E.18)

x	4	5	6
$P(x)$	0.103	0.051	0.005

(5.E.19)

a. What number of customers does Shylock most often see in the bank the moment he enters?
b. What number of customers waiting in line does Shylock most often see the moment he enters?
c. What is the average number of customers who are waiting in line the moment Shylock enters?

27. The owner of a proposed outdoor theater must decide whether to include a cover that will allow shows to be performed in all weather conditions. Based on projected audience sizes and weather conditions, the probability distribution for the revenue X per night if the cover is not installed is

Weather	x	$P(x)$
Clear	\$3,000	0.61
Threatening	\$2,800	0.17
LightRain	\$1,975	0.11
Show − cancelling rain	\$0	0.11

(5.E.20)

The additional cost of the cover is \$410,000. The owner will have it built if this cost can be recovered from the increased revenue the cover affords in the first ten 90-night seasons.

a. Compute the mean revenue per night if the cover is not installed.
b. Use the answer to (a) to compute the projected total revenue per 90-night season if the cover is not installed.
c. Compute the projected total revenue per season when the cover is in place. To do so assume that if the cover were in place the revenue each night of the season would be the same as the revenue on a clear night.
d. Using the answers to (b) and (c), decide whether or not the additional cost of the installation of the cover will be recovered from the increased revenue over the first ten years. Will the owner have the cover installed?

Answers

1. a. no: the sum of the probabilities exceeds 1
   b. no: a negative probability
   c. no: the sum of the probabilities is less than 1
2.
3. a. 0.4
   b. 0.1
   c. 0.9
   d. 79.15
   e. $\sigma^2 = 1.5275$
   f. $\sigma = 1.2359$
4.
5. a. 0.6528
   b. 0.7153
   c. $\mu = 7.8333$
   d. $\sigma^2 = 5.4866$
   e. $\sigma = 2.3424$
6.
7. a. 0.79
   b. 0.60

c. $\mu = 5.8$, $\sigma = 1.2570$

8.

9.

x	0	1	2	3
$P(x)$	1/8	3/8	3/8	1/8

(5.E.21)

10.

11. a.

x	-1	999	499	99
$P(x)$	$\frac{4987}{5000}$	$\frac{1}{5000}$	$\frac{2}{5000}$	$\frac{10}{5000}$

(5.E.22)

b. -0.4

c. 17.8785

12.

13. 136

14.

15. a.

x	C	C	$-150,000$
$P(x)$	0.9825		0.0175

(5.E.23)

b. $C - 2625$

c. $C \geq 2625$

d. $C \geq 2875$

16.

17. a.

x	-1	1
$P(x)$	$\frac{20}{38}$	$\frac{18}{38}$

(5.E.24)

b. $E(X) = -0.0526$. In many bets the bettor sustains an average loss of about 5.25 cents per bet.

c. 0.9986

18.

19. a. 43.54

b. 1.2046

20.

21. 101.02

22.

23. a.

x	0	1	2	3	4	5
$P(x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

(5.E.25)

b. 1.9444

c. 1.4326

24.

25. a.

x	0	1	2
$P(x)$	0.902	0.096	0.002

(5.E.26)

b. 0.902

26.

27. a. 2523.25

b. 227,092.5

c. 270,000

d. The owner will install the cover.

4.3: The Binomial Distribution

Basic

1. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of dots on the top face of fair die that is rolled.
   b. X is the number of hearts in a five-card hand drawn (without replacement) from a well-shuffled ordinary deck.
   c. X is the number of defective parts in a sample of ten randomly selected parts coming from a manufacturing process in which 0.02% of all parts are defective.
   d. X is the number of times the number of dots on the top face of a fair die is even in six rolls of the die.
   e. X is the number of dice that show an even number of dots on the top face when six dice are rolled at once.
2. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of black marbles in a sample of 5 marbles drawn randomly and without replacement from a box that contains 25 white marbles and 15 black marbles.
   b. X is the number of black marbles in a sample of 5 marbles drawn randomly and with replacement from a box that contains 25 white marbles and 15 black marbles.
   c. X is the number of voters in favor of proposed law in a sample $1,200$ randomly selected voters drawn from the entire electorate of a country in which 35% of the voters favor the law.
   d. X is the number of fish of a particular species, among the next ten landed by a commercial fishing boat, that are more than 13 inches in length, when 17% of all such fish exceed 13 inches in length.
   e. X is the number of coins that match at least one other coin when four coins are tossed at once.
3. X is a binomial random variable with parameters $n = 12$ and $p = 0.82$. Compute the probability indicated.
   a. $P(11)$
   b. $P(9)$
   c. $P(0)$
   d. $P(13)$
4. X is a binomial random variable with parameters $n = 16$ and $p = 0.74$. Compute the probability indicated.
   a. $P(14)$
   b. $P(4)$
   c. $P(0)$
   d. $P(20)$
5. X is a binomial random variable with parameters $n = 5$, $p = 0.5$. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $P(X \leq 3)$
   b. $P(X \geq 3)$
   c. $P(3)$
   d. $P(0)$
   e. $P(5)$
6. X is a binomial random variable with parameters $n = 5$, $p = 0.\overline{3}$. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $P(X \leq 2)$
   b. $P(X \geq 2)$
   c. $P(2)$
   d. $P(0)$
   e. $P(5)$
7. X is a binomial random variable with the parameters shown. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.

a. $n = 10, p = 0.25, P(X \leq 6)$
b. $n = 10, p = 0.75, P(X \leq 6)$
c. $n = 15, p = 0.75, P(X \leq 6)$
d. $n = 15, p = 0.75, P(12)$
e. $n = 15, p = 0.\overline{6}, P(10 \leq X \leq 12)$

8. X is a binomial random variable with the parameters shown. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $n = 5, p = 0.05, P(X \leq 1)$
   b. $n = 5, p = 0.5, P(X \leq 1)$
   c. $n = 10, p = 0.75, P(X \leq 5)$
   d. $n = 10, p = 0.75, P(12)$
   e. $n = 10, p = 0.\overline{6}, P(5 \leq X \leq 8)$
9. X is a binomial random variable with the parameters shown. Use the special formulas to compute its mean μ and standard deviation σ.
   a. $n = 8, p = 0.43$
   b. $n = 47, p = 0.82$
   c. $n = 1200, p = 0.44$
   d. $n = 2100, p = 0.62$
10. X is a binomial random variable with the parameters shown. Use the special formulas to compute its mean μ and standard deviation σ.
    a. $n = 14, p = 0.55$
    b. $n = 83, p = 0.05$
    c. $n = 957, p = 0.35$
    d. $n = 1750, p = 0.79$
11. X is a binomial random variable with the parameters shown. Compute its mean μ and standard deviation σ in two ways, first using the tables in 7.1: Large Sample Estimation of a Population Mean in conjunction with the general formulas $\mu = \sum xP(x)$ and $\sigma = \sqrt{[\sum x^2 P(x)] - \mu^2}$, then using the special formulas $\mu = np$ and $\sigma = \sqrt{npq}$.
    a. $n = 5, p = 0.\overline{3}$
    b. $n = 10, p = 0.75$
12. X is a binomial random variable with the parameters shown. Compute its mean μ and standard deviation σ in two ways, first using the tables in 7.1: Large Sample Estimation of a Population Mean in conjunction with the general formulas $\mu = \sum xP(x)$ and $\sigma = \sqrt{[\sum x^2 P(x)] - \mu^2}$, then using the special formulas $\mu = np$ and $\sigma = \sqrt{npq}$.
    a. $n = 10, p = 0.25$
    b. $n = 15, p = 0.1$
13. X is a binomial random variable with parameters $n = 10$ and $p = 1/3$. Use the cumulative probability distribution for X that is given in 7.1: Large Sample Estimation of a Population Mean to construct the probability distribution of X.
14. X is a binomial random variable with parameters $n = 15$ and $p = 1/2$. Use the cumulative probability distribution for X that is given in 7.1: Large Sample Estimation of a Population Mean to construct the probability distribution of X.
15. In a certain board game a player's turn begins with three rolls of a pair of dice. If the player rolls doubles all three times there is a penalty. The probability of rolling doubles in a single roll of a pair of fair dice is $1/6$. Find the probability of rolling doubles all three times.
16. A coin is bent so that the probability that it lands heads up is $2/3$. The coin is tossed ten times.
    a. Find the probability that it lands heads up at most five times.
    b. Find the probability that it lands heads up more times than it lands tails up.

Applications

17. An English-speaking tourist visits a country in which 30% of the population speaks English. He needs to ask someone directions.

a. Find the probability that the first person he encounters will be able to speak English.
b. The tourist sees four local people standing at a bus stop. Find the probability that at least one of them will be able to speak English.

18. The probability that an egg in a retail package is cracked or broken is 0.025.
    a. Find the probability that a carton of one dozen eggs contains no eggs that are either cracked or broken.
    b. Find the probability that a carton of one dozen eggs has (i) at least one that is either cracked or broken; (ii) at least two that are cracked or broken.
    c. Find the average number of cracked or broken eggs in one dozen cartons.
19. An appliance store sells 20 refrigerators each week. Ten percent of all purchasers of a refrigerator buy an extended warranty. Let X denote the number of the next 20 purchasers who do so.
    a. Verify that X satisfies the conditions for a binomial random variable, and find n and p.
    b. Find the probability that X is zero.
    c. Find the probability that X is two, three, or four.
    d. Find the probability that X is at least five.
20. Adverse growing conditions have caused 5% of grapefruit grown in a certain region to be of inferior quality. Grapefruit are sold by the dozen.
    a. Find the average number of inferior quality grapefruit per box of a dozen.
    b. A box that contains two or more grapefruit of inferior quality will cause a strong adverse customer reaction. Find the probability that a box of one dozen grapefruit will contain two or more grapefruit of inferior quality.
21. The probability that a 7-ounce skein of a discount worsted weight knitting yarn contains a knot is 0.25. Goneril buys ten skeins to crochet an afghan.
    a. Find the probability that (i) none of the ten skeins will contain a knot; (ii) at most one will.
    b. Find the expected number of skeins that contain knots.
    c. Find the most likely number of skeins that contain knots.
22. One-third of all patients who undergo a non-invasive but unpleasant medical test require a sedative. A laboratory performs 20 such tests daily. Let X denote the number of patients on any given day who require a sedative.
    a. Verify that X satisfies the conditions for a binomial random variable, and find n and p.
    b. Find the probability that on any given day between five and nine patients will require a sedative (include five and nine).
    c. Find the average number of patients each day who require a sedative.
    d. Using the cumulative probability distribution for X in 7.1: Large Sample Estimation of a Population Mean find the minimum number x_{min} of doses of the sedative that should be on hand at the start of the day so that there is a 99% chance that the laboratory will not run out.
23. About 2% of alumni give money upon receiving a solicitation from the college or university from which they graduated. Find the average number monetary gifts a college can expect from every $2,000$ solicitations it sends.
24. Of all college students who are eligible to give blood, about 18% do so on a regular basis. Each month a local blood bank sends an appeal to give blood to 250 randomly selected students. Find the average number of appeals in such mailings that are made to students who already give blood.
25. About 12% of all individuals write with their left hands. A class of 130 students meets in a classroom with 130 individual desks, exactly 14 of which are constructed for people who write with their left hands. Find the probability that exactly 14 of the students enrolled in the class write with their left hands.
26. A traveling salesman makes a sale on 65% of his calls on regular customers. He makes four sales calls each day.
    a. Construct the probability distribution of X, the number of sales made each day.
    b. Find the probability that, on a randomly selected day, the salesman will make a sale.
    c. Assuming that the salesman makes 20 sales calls per week, find the mean and standard deviation of the number of sales made *per week*.
27. A corporation has advertised heavily to try to insure that over half the adult population recognizes the brand name of its products. In a random sample of 20 adults, 14 recognized its brand name. What is the probability that 14 or more people in such a sample would recognize its brand name if the actual proportion p of all adults who recognize the brand name were only 0.50?

Additional Exercises

28. When dropped on a hard surface a thumbtack lands with its sharp point touching the surface with probability $2/3$; it lands with its sharp point directed up into the air with probability $1/3$. The tack is dropped and its landing position observed 15 times.
    a. Find the probability that it lands with its point in the air at least 7 times.
    b. If the experiment of dropping the tack 15 times is done repeatedly, what is the average number of times it lands with its point in the air?
29. A professional proofreader has a 98% chance of detecting an error in a piece of written work (other than misspellings, double words, and similar errors that are machine detected). A work contains four errors.
    a. Find the probability that the proofreader will miss at least one of them.
    b. Show that two such proofreaders working independently have a 99.96% chance of detecting an error in a piece of written work.
    c. Find the probability that two such proofreaders working independently will miss at least one error in a work that contains four errors.
30. A multiple choice exam has 20 questions; there are four choices for each question.
    a. A student guesses the answer to every question. Find the chance that he guesses correctly between four and seven times.
    b. Find the minimum score the instructor can set so that the probability that a student will pass just by guessing is 20% or less.
31. In spite of the requirement that all dogs boarded in a kennel be inoculated, the chance that a healthy dog boarded in a clean, well-ventilated kennel will develop kennel cough from a carrier is 0.008.
    a. If a carrier (not known to be such, of course) is boarded with three other dogs, what is the probability that at least one of the three healthy dogs will develop kennel cough?
    b. If a carrier is boarded with four other dogs, what is the probability that at least one of the four healthy dogs will develop kennel cough?
    c. The pattern evident from parts (a) and (b) is that if $K+1$ dogs are boarded together, one a carrier and K healthy dogs, then the probability that at least one of the healthy dogs will develop kennel cough is $P(X \geq 1) = 1-(0.992)^K$, where X is the binomial random variable that counts the number of healthy dogs that develop the condition. Experiment with different values of K in this formula to find the maximum number $K+1$ of dogs that a kennel owner can board together so that if one of the dogs has the condition, the chance that another dog will be infected is less than 0.05.
32. Investigators need to determine which of 600 adults have a medical condition that affects 2% of the adult population. A blood sample is taken from each of the individuals.
    a. Show that the expected number of diseased individuals in the group of 600 is 12 individuals.
    b. Instead of testing all 600 blood samples to find the expected 12 diseased individuals, investigators group the samples into 60 groups of 10 each, mix a little of the blood from each of the 10 samples in each group, and test each of the 60 mixtures. Show that the probability that any such mixture will contain the blood of at least one diseased person, hence test positive, is about 0.18.
    c. Based on the result in (b), show that the expected number of mixtures that test positive is about 11. (Supposing that indeed 11 of the 60 mixtures test positive, then we know that none of the 490 persons whose blood was in the remaining 49 samples that tested negative has the disease. We have eliminated 490 persons from our search while performing only 60 tests.)

Answers

1. a. not binomial; not success/failure.
   b. not binomial; trials are not independent.
   c. binomial; $n = 10, p = 0.0002$
   d. binomial; $n = 6, p = 0.5$
   e. binomial; $n = 6, p = 0.5$
2.
3. a. 0.2434
   b. 0.2151
   c. $0.18^{12} \approx 0$

d. 0

4.

5. a. 0.8125
   b. 0.5000
   c. 0.3125
   d. 0.0313
   e. 0.0312

6.

7. a. 0.9965
   b. 0.2241
   c. 0.0042
   d. 0.2252
   e. 0.5390

8.

9. a. $\mu = 3.44, \sigma = 1.4003$
   b. $\mu = 38.54, \sigma = 2.6339$
   c. $\mu = 528, \sigma = 17.1953$
   d. $\mu = 1302, \sigma = 22.2432$

10.

11. a. $\mu = 1.6667, \sigma = 1.0541$
    b. $\mu = 7.5, \sigma = 1.3693$

12.

13.

x	0	1	2	3
$P(x)$	0.0173	0.0867	0.1951	0.2602

(5.E.27)

x	4	5	6	7
$P(x)$	0.2276	0.1365	0.0569	0.0163

(5.E.28)

x	8	9	10
$P(x)$	0.0030	0.0004	0.0000

(5.E.29)

14.

15. 0.0046

16.

17. a. 0.3
    b. 0.7599

18.

19. a. $n = 20, p = 0.1$
    b. 0.1216
    c. 0.5651
    d. 0.0432

20.

21. a. 0.0563 and 0.2440
    b. 2.5
    c. 2

22.

23. 40

24.

25. **0.1019**
26.
27. **0.0577**
28.
29. a. **0.0776**
    b. **0.9996**
    c. **0.0016**
30.
31. a. **0.0238**
    b. **0.0316**
    c. **6**

Contributor

- Anonymous

This page titled 5.E: Discrete Random Variables (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **4.E: Discrete Random Variables (Exercises)** has no license indicated.

CHAPTER OVERVIEW

6: Continuous Random Variables

A random variable is called *continuous* if its set of possible values contains a whole interval of decimal numbers. In this chapter we investigate such random variables.

6.1: The Standard Normal Distribution

6.1.1: Continuous Random Variables

6.1.2: The Standard Normal Distribution

6.2: The General Normal Distribution

6.3: The Central Limit Theorem for Sample Means

6.3E: The Central Limit Theorem for Sample Means (Exercises)

This page titled 6: Continuous Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

6.1: The Standard Normal Distribution

6.1: The Standard Normal Distribution is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.1.1: Continuous Random Variables

Learning Objectives

- To learn the concept of the probability distribution of a continuous random variable, and how it is used to compute probabilities.
- To learn basic facts about the family of normally distributed random variables.

The Probability Distribution of a Continuous Random Variable

For a discrete random variable X the probability that X assumes one of its possible values on a single trial of the experiment makes good sense. This is not the case for a continuous random variable. For example, suppose X denotes the length of time a commuter just arriving at a bus stop has to wait for the next bus. If buses run every 30 minutes without fail, then the set of possible values of X is the interval denoted $[0, 30]$, the set of all decimal numbers between 0 and 30. But although the number 7.211916 is a possible value of X, there is little or no meaning to the concept of the probability that the commuter will wait precisely 7.211916 minutes for the next bus. If anything the probability should be zero, since if we could meaningfully measure the waiting time to the nearest millionth of a minute it is practically inconceivable that we would ever get exactly 7.211916 minutes. More meaningful questions are those of the form: What is the probability that the commuter's waiting time is less than 10 minutes, or is between 5 and 10 minutes? In other words, with continuous random variables one is concerned not with the event that the variable assumes a single particular value, but with the event that the random variable assumes a value in a particular interval.

Definition: density function

The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a function $f(x)$, called a density function, in the following way: the probability that X assumes a value in the interval $[a, b]$ is equal to the area of the region that is bounded above by the graph of the equation $y = f(x)$, bounded below by the x-axis, and bounded on the left and right by the vertical lines through a and b, as illustrated in Figure $6.1.1.1$.

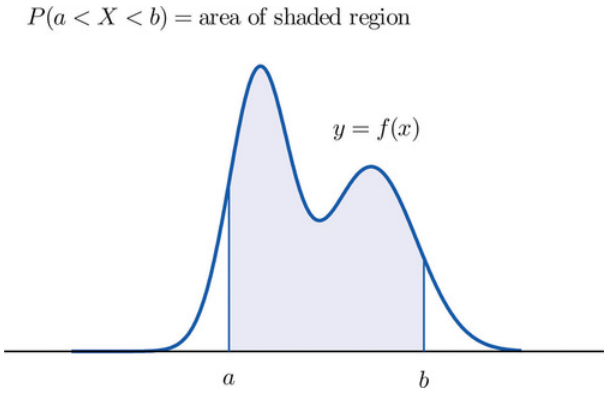


Figure $6.1.1.1$: Probability Given as Area of a Region under a Curve

This definition can be understood as a natural outgrowth of the discussion in Section 2.1.3. There we saw that if we have in view a population (or a very large sample) and make measurements with greater and greater precision, then as the bars in the relative frequency histogram become exceedingly fine their vertical sides merge and disappear, and what is left is just the curve formed by their tops, as shown in Figure 2.1.5. Moreover the total area under the curve is 1, and the proportion of the population with measurements between two numbers a and b is the area under the curve and between a and b, as shown in Figure 2.1.6. If we think of X as a measurement to infinite precision arising from the selection of any one member of the population at random, then $P(a < X < b)$is simply the proportion of the population with measurements between a and b, the curve in the relative frequency histogram is the density function for X, and we arrive at the definition just above.

- Every density function $f(x)$ must satisfy the following two conditions:
- For all numbers x, $f(x) \geq 0$, so that the graph of $y = f(x)$ never drops below the x-axis.
- The area of the region under the graph of $y = f(x)$ and above the x-axis is 1.

Because the area of a line segment is 0, the definition of the probability distribution of a continuous random variable implies that for any particular decimal number, say a, the probability that X assumes the exact value a is 0. This property implies that whether or not the endpoints of an interval are included makes no difference concerning the probability of the interval.

For any continuous random variable X:

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$$

Example 6.1.1.1

A random variable X has the uniform distribution on the interval $[0,1]$: the density function is $f(x) = 1$ if x is between 0 and 1 and $f(x) = 0$ for all other values of x, as shown in Figure 6.1.1.2.

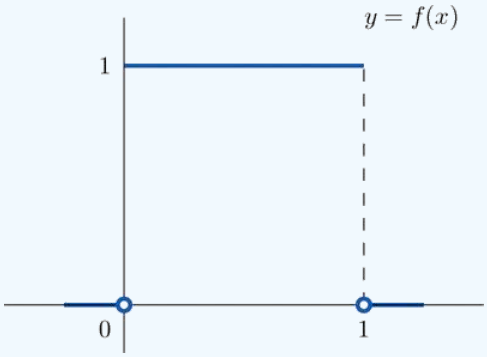


Figure 6.1.1.2: Uniform Distribution on [0,1].

1. Find $P(X > 0.75)$, the probability that X assumes a value greater than 0.75.
2. Find $P(X \le 0.2)$, the probability that X assumes a value less than or equal to 0.2.
3. Find $P(0.4 < X < 0.7)$, the probability that X assumes a value between 0.4 and 0.7.

Solution

1. $P(X > 0.75)$ is the area of the rectangle of height 1 and base length $1 - 0.75 = 0.25$, hence is $base \times height = (0.25) \cdot (1) = 0.25$. See Figure 6.1.1.3a.
2. $P(X \le 0.2)$ is the area of the rectangle of height 1 and base length $0.2 - 0 = 0.2$, hence is $base \times height = (0.2) \cdot (1) = 0.2$. See Figure 6.1.1.3b.
3. $P(0.4 < X < 0.7)$ is the area of the rectangle of height 1 and length $0.7 - 0.4 = 0.3$, hence is $base \times height = (0.3) \cdot (1) = 0.3$. See Figure 6.1.1.3c.

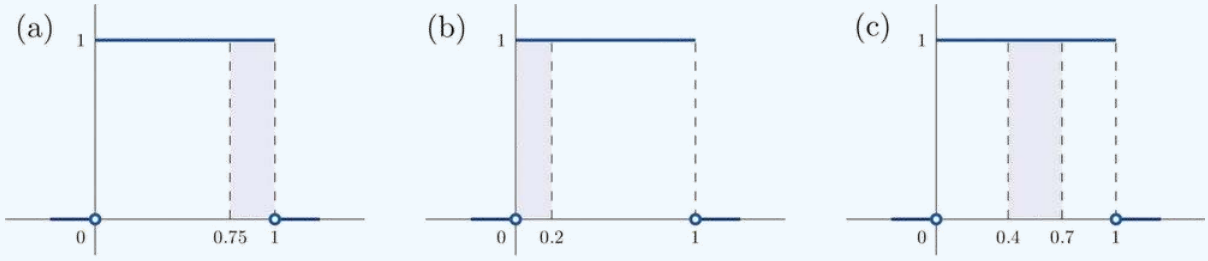


Figure 6.1.1.3: Probabilities from the Uniform Distribution on [0,1]

Example 6.1.1.2

A man arrives at a bus stop at a random time (that is, with no regard for the scheduled service) to catch the next bus. Buses run every 30 minutes without fail, hence the next bus will come any time during the next 30 minutes with evenly distributed probability (a uniform distribution). Find the probability that a bus will come within the next 10 minutes.

Solution

The graph of the density function is a horizontal line above the interval from 0 to 30 and is the x-axis everywhere else. Since the total area under the curve must be 1, the height of the horizontal line is $1/30$ (Figure 6.1.1.4). The probability sought is $P(0$

$\leq X \leq 10$).By definition, this probability is the area of the rectangular region bounded above by the horizontal line $f(x) = 1/30$, bounded below by the x-axis, bounded on the left by the vertical line at 0 (the y-axis), and bounded on the right by the vertical line at 10. This is the shaded region in Figure 6.1.1.4. Its area is the base of the rectangle times its height, $(10) \cdot (1/30) = 1/3$. Thus $P(0 \leq X \leq 10) = 1/3$.

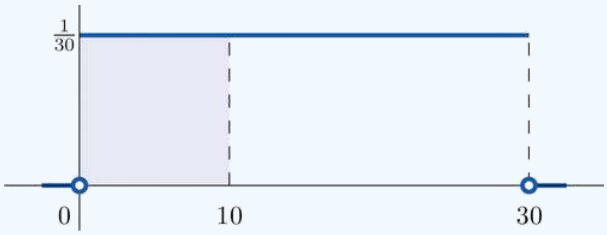


Figure 6.1.1.4: Probability of Waiting At Most 10 Minutes for a Bus

Normal Distributions

Most people have heard of the "bell curve." It is the graph of a specific density function $f(x)$ that describes the behavior of continuous random variables as different as the heights of human beings, the amount of a product in a container that was filled by a high-speed packing machine, or the velocities of molecules in a gas. The formula for $f(x)$ contains two parameters μ and σ that can be assigned any specific numerical values, so long as σ is positive. We will not need to know the formula for $f(x)$, but for those who are interested it is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\mu - x)^2/\sigma^2}$$

where $\pi \approx 3.14159$ and $e \approx 2.71828$ is the base of the natural logarithms.

Each different choice of specific numerical values for the pair μ and σ gives a different bell curve. The value of μ determines the location of the curve, as shown in Figure 6.1.1.5. In each case the curve is symmetric about μ.

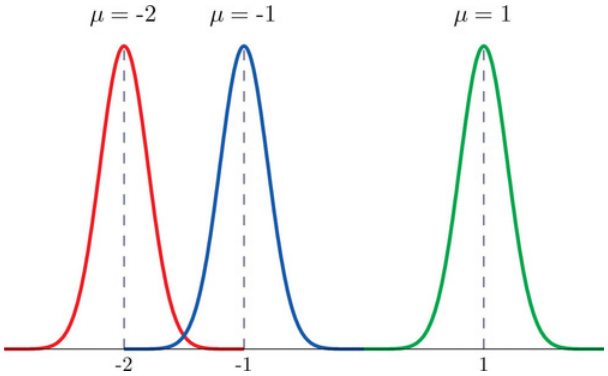


Figure 6.1.1.5: Bell Curves with σ = 0.25 and Different Values of μ

The value of σ determines whether the bell curve is tall and thin or short and squat, subject always to the condition that the total area under the curve be equal to 1. This is shown in Figure 6.1.1.6, where we have arbitrarily chosen to center the curves at $\mu = 6$.

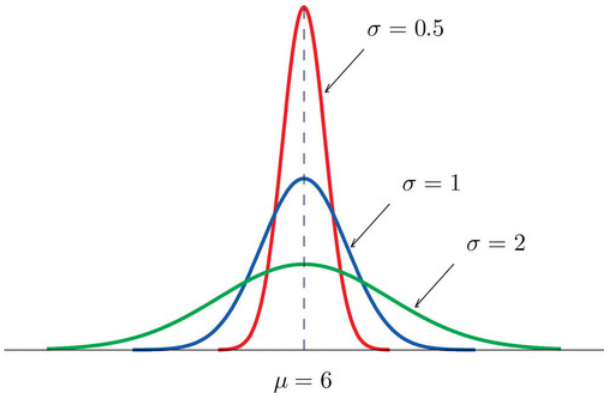


Figure **6.1.1.6**: Bell Curves with $\mu = 6$ and Different Values of σ.

Definition: normal distribution

The probability distribution corresponding to the density function for the bell curve with parameters μ and σ is called the normal distribution with mean μ and standard deviation σ.

Definition: normally distributed random variable

A continuous random variable whose probabilities are described by the normal distribution with mean μ and standard deviation σ is called a normally distributed random variable, or a normal random variable for short, with mean μ and standard deviation σ.

Figure **6.1.1.7** shows the density function that determines the normal distribution with mean μ and standard deviation σ. We repeat an important fact about this curve: **The density curve for the normal distribution is symmetric about the mean.**

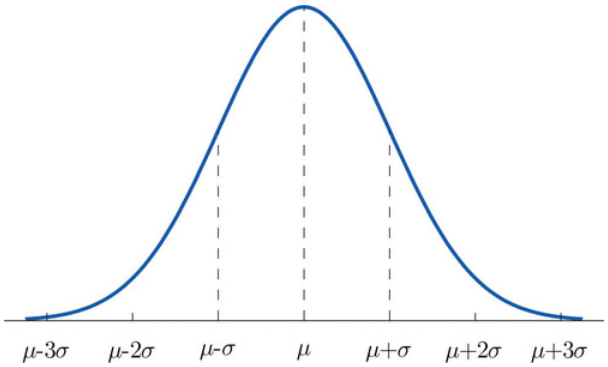


Figure **6.1.1.7**: Density Function for a Normally Distributed Random Variable with Mean μ and Standard Deviation σ

Example 6.1.1.3

Heights of **25**-year-old men in a certain region have mean **69.75** inches and standard deviation **2.59** inches. These heights are approximately normally distributed. Thus the height X of a randomly selected **25**-year-old man is a normal random variable with mean $\mu = 69.75$ and standard deviation $\sigma = 2.59$. Sketch a qualitatively accurate graph of the density function for X. Find the probability that a randomly selected **25**-year-old man is more than **69.75** inches tall.

Solution

The distribution of heights looks like the bell curve in Figure **6.1.1.8**. The important point is that it is centered at its mean, **69.75**, and is symmetric about the mean.

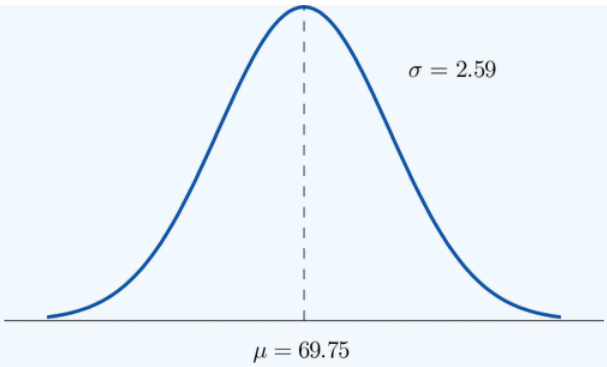


Figure **6.1.1.8**: Density Function for Heights of **25**-Year-Old Men

Since the total area under the curve is **1**, by symmetry the area to the right of **69.75** is half the total, or **0.5**. But this area is precisely the probability $P(X > 69.75)$, the probability that a randomly selected **25**-year-old man is more than **69.75** inches tall. We will learn how to compute other probabilities in the next two sections.

Key Takeaway

- For a continuous random variable X the only probabilities that are computed are those of X taking a value in a specified interval.
- The probability that X take a value in a particular interval is the same whether or not the endpoints of the interval are included.
- The probability $P(a < X < b)$, that X take a value in the interval from a to b, is the area of the region between the vertical lines through a and b, above the x-axis, and below the graph of a function $f(x)$ called the density function.
- A normally distributed random variable is one whose density function is a bell curve.
- Every bell curve is symmetric about its mean and lies everywhere above the x-axis, which it approaches asymptotically (arbitrarily closely without touching).

This page titled 6.1.1: Continuous Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **5.1: Continuous Random Variables** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

6.1.2: The Standard Normal Distribution

Learning Objectives

- To learn what a standard normal random variable is.
- To learn how to compute probabilities related to a standard normal random variable.

Definition: standard normal random variable

A *standard normal random variable* is a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$. It will always be denoted by the letter Z.

The density function for a standard normal random variable is shown in Figure 6.1.2.1.

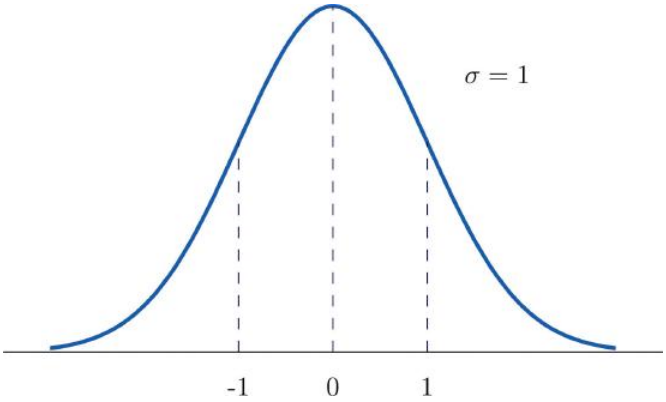


Figure 6.1.2.1: Density Curve for a Standard Normal Random Variable

To compute probabilities for Z we will not work with its density function directly but instead read probabilities out of Figure 6.1.2.2. The tables are tables of *cumulative* probabilities; their entries are probabilities of the form $P(Z < z)$. The use of the tables will be explained by the following series of examples.

Example 6.1.2.1

Find the probabilities indicated, where as always Z denotes a standard normal random variable.

1. $P(Z < 1.48)$.
2. $P(Z < -0.25)$.

Solution

1. Figure 6.1.2.3 shows how this probability is read directly from the table without any computation required. The digits in the ones and tenths places of 1.48, namely 1.4, are used to select the appropriate row of the table; the hundredths part of 1.48, namely 0.08, is used to select the appropriate column of the table. The four decimal place number in the interior of the table that lies in the intersection of the row and column selected, 0.9306, is the probability sought:

$$P(Z < 1.48) = 0.9306$$

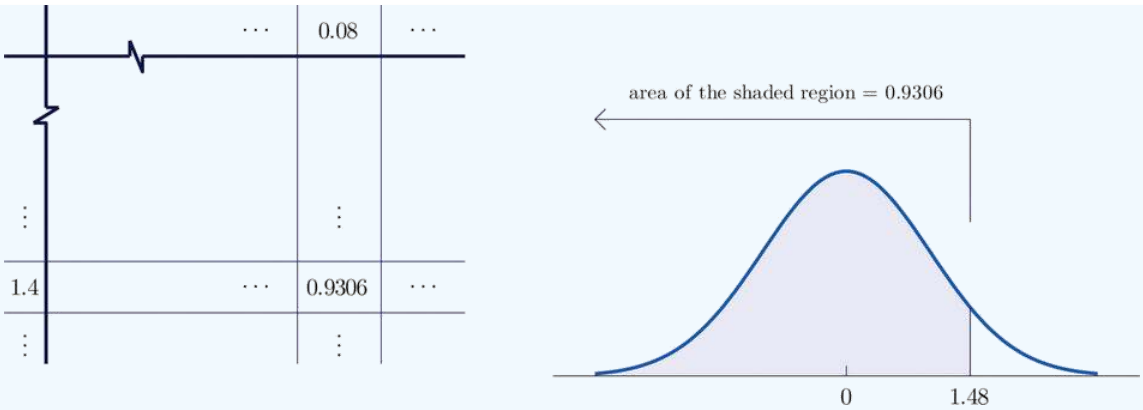


Figure 6.1.2.3: *Computing Probabilities Using the Cumulative Table*

2. The minus sign in -0.25 makes no difference in the procedure; the table is used in exactly the same way as in part (a): the probability sought is the number that is in the intersection of the row with heading -0.2 and the column with heading 0.05, the number 0.4013. Thus $P(Z < -0.25) = 0.4013$.

Example 6.1.2.2

Find the probabilities indicated.

1. $P(Z > 1.60)$.
2. $P(Z > -1.02)$.

Solution

1. Because the events $Z > 1.60$ and $Z \leq 1.60$ are complements, the *Probability Rule for Complements* implies that

$$P(Z > 1.60) = 1 - P(Z \leq 1.60)$$

Since inclusion of the endpoint makes no difference for the continuous random variable Z, $P(Z \leq 1.60) = P(Z < 1.60)$, which we know how to find from the table in Figure 6.1.2.2. The number in the row with heading 1.6 and in the column with heading 0.00 is 0.9452. Thus $P(Z < 1.60) = 0.9452$ so

$$P(Z > 1.60) = 1 - P(Z \leq 1.60) = 1 - 0.9452 = 0.0548$$

Figure 6.1.2.4 illustrates the ideas geometrically. Since the total area under the curve is 1 and the area of the region to the left of 1.60 is (from the table) 0.9452, the area of the region to the right of 1.60 must be $1 - 0.9452 = 0.0548$.

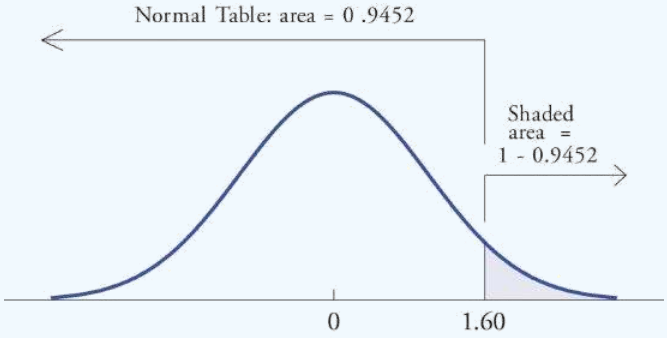


Figure 6.1.2.4: Computing a Probability for a Right Half-Line

2. The minus sign in -1.02 makes no difference in the procedure; the table is used in exactly the same way as in part (a). The number in the intersection of the row with heading -1.0 and the column with heading 0.02 is 0.1539. This means that $P(Z < -1.02) = P(Z \leq -1.02) = 0.1539$. Hence

$$P(Z > -1.02) = P(Z \leq -1.02) = 1 - 0.1539 = 0.8461$$

Example 6.1.2.3

Find the probabilities indicated.

1. $P(0.5 < Z < 1.57)$.
2. $P(-2.55 < Z < 0.09)$.

Solution

1. Figure **6.1.2.5** illustrates the ideas involved for intervals of this type. First look up the areas in the table that correspond to the numbers **0.5** (which we think of as **0.50** to use the table) and **1.57**. We obtain **0.6915** and **0.9418**, respectively. From the figure it is apparent that we must take the difference of these two numbers to obtain the probability desired. In symbols,

$$P(0.5 < Z < 1.57) = P(Z < 1.57) - P(Z < 0.50) = 0.9418 - 0.6915 = 0.2503$$

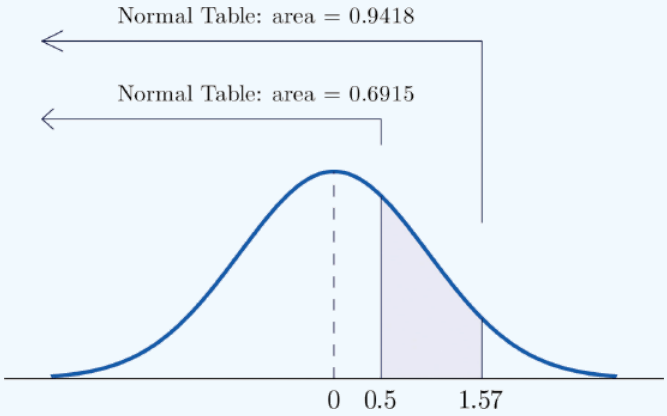


Figure **6.1.2.5**: Computing a Probability for an Interval of Finite Length

2. The procedure for finding the probability that Z takes a value in a finite interval whose endpoints have opposite signs is exactly the same procedure used in part (a), and is illustrated in Figure **6.1.2.6** "Computing a Probability for an Interval of Finite Length". In symbols the computation is

$$P(-2.55 < Z < 0.09) = P(Z < 0.09) - P(Z < -2.55) = 0.5359 - 0.0054 = 0.5305$$

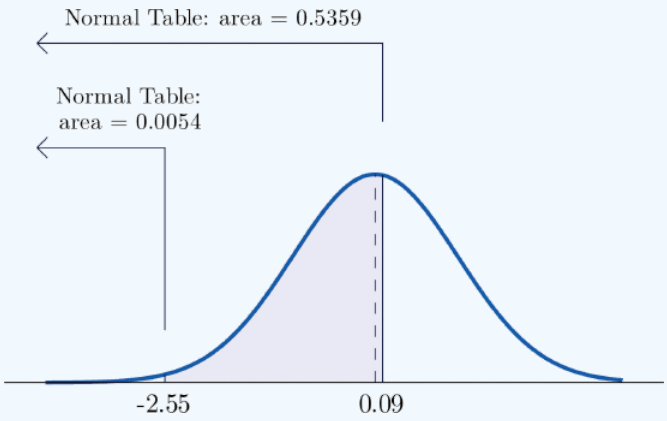


Figure **6.1.2.6**: Computing a Probability for an Interval of Finite Length

The next example shows what to do if the value of Z that we want to look up in the table is not present there.

Example 6.1.2.4

Find the probabilities indicated.

1. $P(1.13 < Z < 4.16)$.
2. $P(-5.22 < Z < 2.15)$.

Solution

1. We attempt to compute the probability exactly as in Example 6.1.2.3 by looking up the numbers 1.13 and 4.16 in the table. We obtain the value 0.8708 for the area of the region under the density curve to left of 1.13 without any problem, but when we go to look up the number 4.16 in the table, it is not there. We can see from the last row of numbers in the table that the area to the left of 4.16 must be so close to 1 that to four decimal places it rounds to 1.0000. Therefore

$$P(1.13 < Z < 4.16) = 1.0000 - 0.8708 = 0.1292$$

2. Similarly, here we can read directly from the table that the area under the density curve and to the left of 2.15 is 0.9842, but −5.22 is too far to the left on the number line to be in the table. We can see from the first line of the table that the area to the left of −5.22 must be so close to 0 that to four decimal places it rounds to 0.0000. Therefore

$$P(-5.22 < Z < 2.15) = 0.9842 - 0.0000 = 0.9842$$

The final example of this section explains the origin of the proportions given in the Empirical Rule.

Example 6.1.2.5

Find the probabilities indicated.

1. $P(-1 < Z < 1)$.
2. $P(-2 < Z < 2)$.
3. $P(-3 < Z < 3)$.

Solution

1. Using the table as was done in Example 6.1.2.3 we obtain

$$P(-1 < Z < 1) = 0.8413 - 0.1587 = 0.6826$$

   Since Z has mean 0 and standard deviation 1, for Z to take a value between −1 and 1 means that Z takes a value that is within one standard deviation of the mean. Our computation shows that the probability that this happens is about 0.68, the proportion given by the Empirical Rule for histograms that are mound shaped and symmetrical, like the bell curve.
2. Using the table in the same way,

$$P(-2 < Z < 2) = 0.9772 - 0.0228 = 0.9544$$

   This corresponds to the proportion 0.95 for data within two standard deviations of the mean.
3. Similarly,

$$P(-3 < Z < 3) = 0.9987 - 0.0013 = 0.9974$$

   which corresponds to the proportion 0.997 for data within three standard deviations of the mean.

Key takeaway

- A standard normal random variable Z is a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$.
- Probabilities for a standard normal random variable are computed using Figure 6.1.2.2.

This page titled 6.1.2: The Standard Normal Distribution is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **5.2: The Standard Normal Distribution** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

6.2: The General Normal Distribution

Learning Objectives

- To learn how to compute probabilities related to any normal random variable.

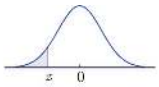
If X is any normally distributed normal random variable then Figure **6.2.1** can also be used to compute a probability of the form $P(a < X < b)$ by means of the following equality.

equality

If X is a normally distributed random variable with mean μ and standard deviation σ, then

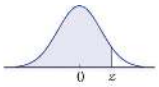
$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

where Z denotes a standard normal random variable. a can be any decimal number or $-\infty$; b can be any decimal number or ∞.



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8304	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure **6.2.1**: Cumulative Normal Probability

The new endpoints $\frac{(a-\mu)}{\sigma}$ and $\frac{(b-\mu)}{\sigma}$ are the z-scores of a and b as defined in Chapter 2.

Figure **6.2.2** illustrates the meaning of the equality geometrically: the two shaded regions, one under the density curve for X and the other under the density curve for Z, have the same area. Instead of drawing both bell curves, though, we will always draw a single generic bell-shaped curve with both an x-axis and a z-axis below it.

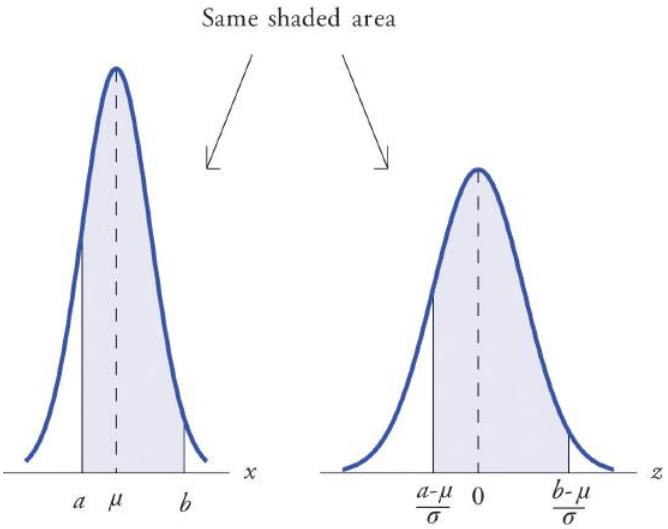


Figure **6.2.2**: Probability for an Interval of Finite Length

Example 6.2.1

Let X be a normal random variable with mean $\mu = 10$ and standard deviation $\sigma = 2.5$. Compute the following probabilities.

1. $P(X < 14)$.
2. $P(8 < X < 14)$.

Solution

1. See Figure 6.2.3 "Probability Computation for a General Normal Random Variable".

$$\begin{aligned} P(X < 14) &= P\left(Z < \frac{14-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{14-10}{2.5}\right) \\ &= P(Z < 1.60) \\ &= 0.9452 \end{aligned}$$

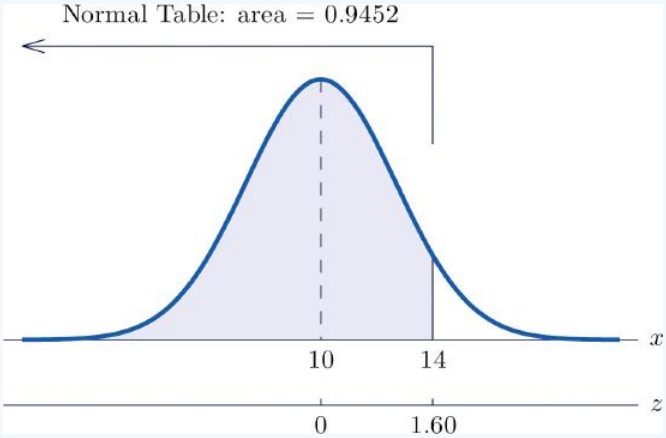


Figure 6.2.3: *Probability Computation for a General Normal Random Variable*

2. See Figure 6.2.4 "Probability Computation for a General Normal Random Variable".

$$\begin{aligned} P(8 < X < 14) &= P\left(\frac{8-10}{2.5} < Z < \frac{14-10}{2.5}\right) \\ &= P(-0.80 < Z < 1.60) \\ &= 0.9452 - 0.2119 \\ &= 0.7333 \end{aligned}$$

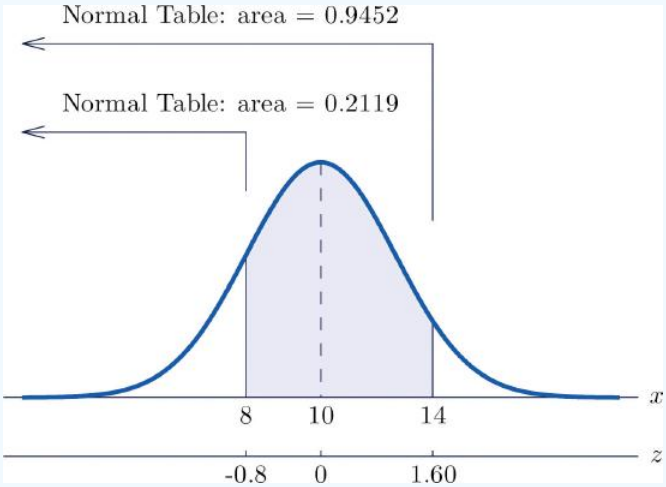


Figure 6.2.4: *Probability Computation for a General Normal Random Variable*

Example 6.2.2

The lifetimes of the tread of a certain automobile tire are normally distributed with mean $37,500$ miles and standard deviation $4,500$ miles. Find the probability that the tread life of a randomly selected tire will be between $30,000$ and $40,000$ miles.

Solution

Let X denote the tread life of a randomly selected tire. To make the numbers easier to work with we will choose thousands of miles as the units. Thus $\mu = 37.5$, $\sigma = 4.5$, and the problem is to compute $P(30 < X < 40)$. Figure 6.2.5 "Probability Computation for Tire Tread Wear" illustrates the following computation:

$$\begin{aligned} P(30 < X < 40) &= P\left(\frac{30-\mu}{\sigma} < Z < \frac{40-\mu}{\sigma}\right) \\ &= P\left(\frac{30-37.5}{4.5} < Z < \frac{40-37.5}{4.5}\right) \\ &= P(-1.67 < Z < 0.56) \\ &= 0.7123 - 0.0475 \\ &= 0.6648 \end{aligned}$$

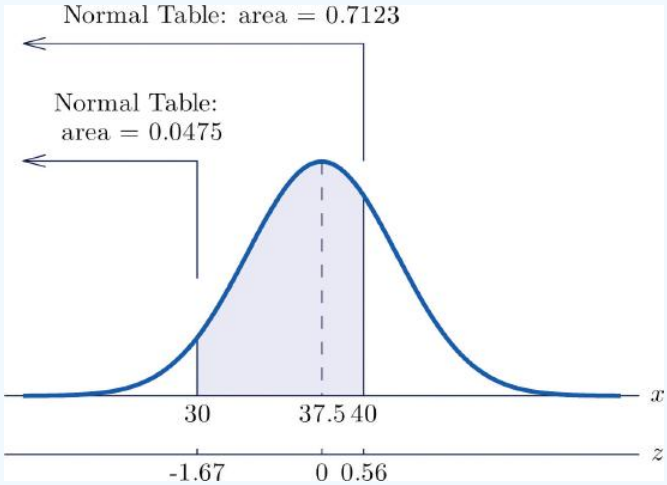


Figure 6.2.5: Probability Computation for Tire Tread Wear

Note that the two z-scores were rounded to two decimal places in order to use Figure 6.2.1 "Cumulative Normal Probability".

Example 6.2.3

Scores on a standardized college entrance examination (*CEE*) are normally distributed with mean 510 and standard deviation 60. A selective university considers for admission only applicants with *CEE* scores over 650. Find percentage of all individuals who took the *CEE* who meet the university's *CEE* requirement for consideration for admission.

Solution

Let X denote the score made on the *CEE* by a randomly selected individual. Then X is normally distributed with mean 510 and standard deviation 60. The probability that X lie in a particular interval is the same as the proportion of all exam scores that lie in that interval. Thus the solution to the problem is $P(X > 650)$, expressed as a percentage. Figure 6.2.6 "Probability Computation for Exam Scores" illustrates the following computation:

$$\begin{aligned} P(X > 650) &= P\left(Z > \frac{650-\mu}{\sigma}\right) \\ &= P\left(Z > \frac{650-510}{60}\right) \\ &= P(Z > 2.33) \\ &= 1 - 0.9901 \\ &= 0.0099 \end{aligned}$$

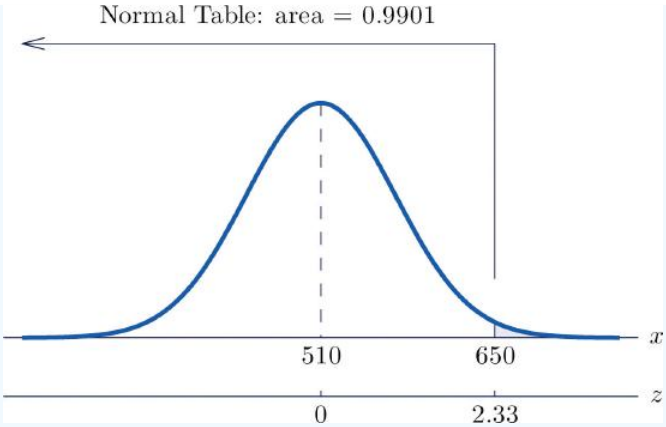


Figure **6.2.6**: *Probability Computation for Exam Scores*

The proportion of all *CEE* scores that exceed **650** is **0.0099**, hence **0.99%** or about **1%** do.

- Probabilities for a general normal random variable are computed using Figure **6.2.1** after converting x-values to z-scores.

This page titled 6.2: The General Normal Distribution is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **5.3: Probability Computations for General Normal Random Variables** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

6.3: The Central Limit Theorem for Sample Means

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

a. $\mu_x =$ the mean of X
b. $\sigma_x =$ the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\bar{X}$ which consists of sample means, tends to be normally distributed and

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right). \tag{6.3.1}$$

The central limit theorem for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and calculating their means, the sample means form their own *normal distribution* (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by, the sample size. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n, the distribution of the random variable $\bar{X}$, which consists of sample means, is called the *sampling distribution of the mean*. The sampling distribution of the mean approaches a normal distribution as n, the sample size, increases.

The random variable $\bar{X}$ has a different z-score associated with it from that of the random variable X. The mean $\bar{x}$ is the value of $\bar{X}$ in one sample.

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)} \tag{6.3.2}$$

- μ_x is the average of both X and $\bar{X}$.
- $\sigma\bar{x} = \frac{\sigma_x}{\sqrt{n}} =$ standard deviation of $\bar{X}$ and is called the standard error of the mean.

Howto: Find probabilities for means on the calculator

2nd DISTR

2:normalcdf

$$\text{normalcdf}\left(\text{lower value of the area, upper value of the area, mean, } \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}\right)$$

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* $= n$

Example 6.3.1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

a. Find the probability that the sample mean is between 85 and 92.
b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

Answer

a.

Let $X =$ one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

Let $\bar{X} =$ the mean of a sample of size 25. Since $\mu_x = 90, \sigma_x = 15$, and $n = 25$,

$$\bar{X} \sim N(90, \frac{15}{\sqrt{25}}).$$

Find $P(85 < x < 92)$. Draw a graph.

$$P(85 < x < 92) = 0.6997$$

The probability that the sample mean is between 85 and 92 is 0.6997.

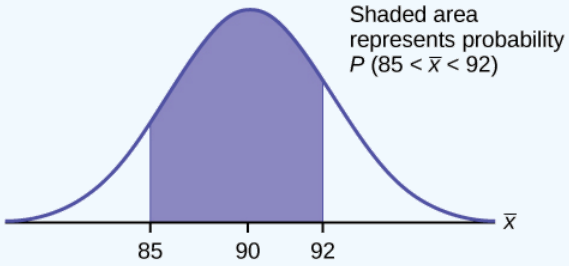


Figure **6.3.1**.

`normalcdf` (lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$)

`normalcdf` $(85, 92, 90, \frac{15}{\sqrt{25}}) = 0.6997$

b.

To find the value that is two standard deviations above the expected value 90, use the formula:

$$\text{value} = \mu_x + (\#\text{ofTSDEVs})\left(\frac{\sigma_x}{\sqrt{n}}\right)$$

$$= 90 + 2\left(\frac{15}{\sqrt{25}}\right) = 96$$

The value that is two standard deviations above the expected value is 96.

The standard error of the mean is

$$\frac{\sigma_x}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3.$$

Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size n.

? Exercise 6.3.1

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size n = 30 are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Answer

$$P(42 < \bar{x} < 50) = \left(42, 50, 45, \frac{8}{\sqrt{30}}\right) = 0.9797$$

Example 6.3.2

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size** $n = 50$ is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Answer

Let $X =$ the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.

Let $\bar{X} =$ the mean time, in hours, it takes to play one soccer match.

If $\mu_x =$ _____, $\sigma_x =$ _____, and $n =$ _____, then $X \sim N$(_____, _____) by the central limit theorem for means.

$\mu_x = 2, \sigma_x = 0.5, n = 50$, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < \bar{x} < 2.3)$. Draw a graph.

$P(1.8 < \bar{x} < 2.3) = 0.9977$

normalcdf $\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

Exercise 6.3.2

The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

Answer

$$P(2 < \bar{x} < 3) = \text{normalcdf}\left(2, 3, 2.5, \frac{0.25}{\sqrt{60}}\right) = 1$$

Calculator SKills

To find percentiles for means on the calculator, follow these steps.

- 2nd DIStR
- 3:invNorm

$$k = \text{invNorm}\left(\text{area to the left of} k, \text{mean}, \frac{\text{standard deviation}}{\sqrt{samplesize}}\right)$$

where:

- k = the kth percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example 6.3.3

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size $n = 100$.

a. What are the mean and standard deviation for the sample mean ages of tablet users?
b. What does the distribution look like?
c. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
d. Find the 95th percentile for the sample mean age (to one decimal place).

Answer

a. Since the sample mean tends to target the population mean, we have $\mu_x = \mu = 34$. The sample standard deviation is given by:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

b. The central limit theorem states that for large sample sizes (n), the sampling distribution will be approximately normal.
c. The probability that the sample mean age is more than 30 is given by:

$$P(\mathrm{X} > 30) = \mathrm{normalcdf}(30, E99, 34, 1.5) = 0.9962$$

d. Let k = the 95th percentile.

$$k = \mathrm{invNorm}\left(0.95, 34, \frac{15}{\sqrt{100}}\right) = 36.5$$

Exercise 6.3.3

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Answer

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

$$P(29 < \bar{x} < 35) = \mathrm{normalcdf}\left(29, 35, 28, \frac{4.8}{\sqrt{100}}\right) = 0.0186 \quad (6.3.3)$$

You can conclude there is approximately a 1.9% chance that your game will be played by men whose mean age is between 29 and 35.

Example 6.3.4

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

a. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
b. What is the standard error of the mean?
c. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
d. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Answer

a. $\mu = \mu = 8.2 \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$

b. This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.

c. Let k = the 90th percentile

$k = \text{invNorm}\left(0.90, 8.2, \frac{1}{\sqrt{60}}\right) = 8.37$. This values indicates that 90 percent of the average app engagement time for table users is less than 8.37 minutes.

d. $P(8 < \bar{x} < 8.5) = \text{normalcdf}\left(8, 8.5, 8.2, \frac{1}{\sqrt{60}}\right) = 0.9293$

? Exercise 6.3.4

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are $n = 34$, $\bar{x} = 16.01$ ounces. If the cans are filled so that $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Answer

We have $P(\bar{x} > 16.01) = \text{normalcdf}\left(16.01, E99, 16, \frac{0.143}{\sqrt{34}}\right) = 0.3417$. Since there is a 34.17% probability that the average sample weight is greater than 16.01 ounces, we should be skeptical of the company's claimed volume. If I am a consumer, I should be glad that I am probably receiving free cola. If I am the manufacturer, I need to determine if my bottling processes are outside of acceptable limits.

Summary

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).

Formula Review

- The Central Limit Theorem for Sample Means:

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

- The Mean $\bar{X}$: σ_x
- Central Limit Theorem for Sample Means z-score and standard error of the mean:

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

- Standard Error of the Mean (Standard Deviation ($\bar{X}$)):

$$\frac{\sigma_x}{\sqrt{n}}$$

Glossary

Average

a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

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, $\bar{X}$, and the sample sum, $\sum X$. If the size (n) of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X \sim N(n\mu, (\sqrt{n})(\sigma))$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean, and the mean of the sample sums will equal n times 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.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, 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 a **standard normal distribution**.

Standard Error of the Mean

the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.

References

1. Baran, Daya. “20 Percent of Americans Have Never Used Email.”WebGuild, 2010. Available online at www.webguild.org/20080519/20-...ver-used-email (accessed May 17, 2013).
2. Data from The Flurry Blog, 2013. Available online at blog.flurry.com (accessed May 17, 2013).
3. Data from the United States Department of Agriculture.

This page titled 6.3: The Central Limit Theorem for Sample Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.3E: The Central Limit Theorem for Sample Means (Exercises)

Use the following information to answer the next six exercises: Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\bar{X}$ be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

? Example 6.3E.1

What is the mean, standard deviation, and sample size?

Answer

mean = 4 hours; standard deviation = 1.2 hours; sample size = 16

Exercise 6.3E.2

Complete the distributions.

1. $X \sim$ _____(_____,_____)
2. $\bar{X} \sim$ _____(_____,_____)

? Example 6.3E.3

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



***Figure* 6.3E.2.**

2. P(_____ $< x <$ _____) = _____

Answer

1. Check student's solution.
2. 3.5, 4.25, 0.2441

Exercise 6.3E.4

Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



***Figure* 6.3E.3.**

2. P(_____) = _____

Example 6.3E.5

What causes the probabilities in Exercise and Exercise to be different?

Answer

The fact that the two distributions are different accounts for the different probabilities.

Exercise 6.3E.6

Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.



***Figure* 6.3E.4.**

a. The 95th Percentile =_____

This page titled 6.3E: The Central Limit Theorem for Sample Means (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

7: Estimation

If we wish to estimate the mean μ of a population for which a census is impractical, say the average height of all 18-year-old men in the country, a reasonable strategy is to take a sample, compute its mean $\bar{x}$, and estimate the unknown number μ by the known number $\bar{x}$. For example, if the average height of 100 randomly selected men aged 18 is 70.6 inches, then we would say that the average height of all 18-year-old men is (at least approximately) 70.6 inches.

Estimating a population parameter by a single number like this is called point estimation; in the case at hand the statistic $\bar{x}$ is a point estimate of the parameter μ. The terminology arises because a single number corresponds to a single point on the number line.

A problem with a point estimate is that it gives no indication of how reliable the estimate is. In contrast, in this chapter we learn about interval estimation. In brief, in the case of estimating a population mean μ we use a formula to compute from the data a number E, called the margin of error of the estimate, and form the interval $[\bar{x}-E,\bar{x}+E]$. We do this in such a way that a certain proportion, say 95%, of all the intervals constructed from sample data by means of this formula contain the unknown parameter μ. Such an interval is called a 95% confidence interval for μ.

Continuing with the example of the average height of 18-year-old men, suppose that the sample of 100 men mentioned above for which $\bar{x}=70.6$ inches also had sample standard deviation $s = 1.7$ inches. It then turns out that $E = 0.33$ and we would state that we are 95% confident that the average height of all 18-year-old men is in the interval formed by 70.6 ± 0.33 inches, that is, the average is between 70.27 and 70.93 inches. If the sample statistics had come from a smaller sample, say a sample of 50 men, the lower reliability would show up in the 95% confidence interval being longer, hence less precise in its estimate. In this example the 95% confidence interval for the same sample statistics but with $n = 50$ is 70.6 ± 0.47 inches, or from 70.13 to 71.07 inches.

7.1: Estimation of a Population Proportion

7.2: Estimation of a Population Mean

7.2.1: Large Sample Estimation of a Population Mean

7.2.2: Small Sample Estimation of a Population Mean

7.3: Sample Size Considerations

7.E: Estimation (Exercises)

This page titled 7: Estimation is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

7.1: Estimation of a Population Proportion

Learning Objectives

- To understand how to apply the formula for a confidence interval for a population proportion.

Since from Section 6.3, we know the mean, standard deviation, and sampling distribution of the sample proportion $\hat{p}$, the ideas of the previous two sections can be applied to produce a confidence interval for a population proportion. Here is the formula.

Large Sample $100(1-\alpha)\%$ Confidence Interval for a Population Proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

A sample is large if the interval $[p-3\sigma_{\hat{p}}, p+3\sigma_{\hat{p}}]$ lies wholly within the interval $[0,1]$.

In actual practice the value of p is not known, hence neither is $\sigma_{\hat{p}}$. In that case we substitute the known quantity $\hat{p}$ for p in making the check; this means checking that the interval

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

lies wholly within the interval $[0,1]$.

Example 7.1.1

To estimate the proportion of students at a large college who are female, a random sample of 120 students is selected. There are 69 female students in the sample. Construct a 90% confidence interval for the proportion of all students at the college who are female.

Solution

The proportion of students in the sample who are female is

$$\hat{p} = 69/120 = 0.575$$

Confidence level 90% means that $\alpha = 1-0.90 = 0.10$ so $\alpha/2 = 0.05$. From the last line of Figure 7.1.6 we obtain $z_{0.05} = 1.645$.

Thus

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.575 \pm 1.645\sqrt{\frac{(0.575)(0.425)}{120}} = 0.575 \pm 0.074$$

One may be 90% confident that the true proportion of all students at the college who are female is contained in the interval $(0.575-0.074, 0.575+0.074) = (0.501, 0.649)$.

Summary

- We have a single formula for a confidence interval for a population proportion, which is valid when the sample is large.
- The condition that a sample be large is not that its size n be at least 30, but that the density function fit inside the interval $[0,1]$.

This page titled 7.1: Estimation of a Population Proportion is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **7.3: Large Sample Estimation of a Population Proportion** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

7.2: Estimation of a Population Mean

7.2: Estimation of a Population Mean is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.2.1: Large Sample Estimation of a Population Mean

Learning Objectives

- To become familiar with the concept of an interval estimate of the population mean.
- To understand how to apply formulas for a confidence interval for a population mean.

The Central Limit Theorem says that, for large samples (samples of size $n \geq 30$), when viewed as a random variable the sample mean $\overline{X}$ is normally distributed with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$. The Empirical Rule says that we must go about two standard deviations from the mean to capture 95% of the values of $\overline{X}$ generated by sample after sample. A more precise distance based on the normality of $\overline{X}$ is 1.960 standard deviations, which is $E = \frac{1.960\sigma}{\sqrt{n}}$.

The key idea in the construction of the 95% confidence interval is this, as illustrated in Figure 7.2.1.1, because in sample after sample 95% of the values of $\overline{X}$ lie in the interval $[\mu - E, \mu + E]$, if we adjoin to each side of the point estimate $x - a$ "wing" of length E, 95% of the intervals formed by the winged dots contain μ. The 95% confidence interval is thus $\bar{x} \pm 1.960\frac{\sigma}{\sqrt{n}}$. For a different level of confidence, say 90% or 99%, the number 1.960 will change, but the idea is the same.

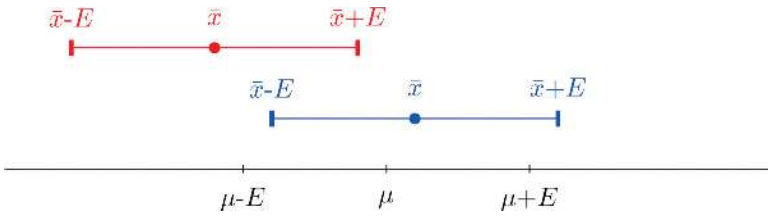


Figure 7.2.1.1: When Winged Dots Capture the Population Mean

Figure 7.2.1.2 shows the intervals generated by a computer simulation of drawing 40 samples from a normally distributed population and constructing the 95% confidence interval for each one. We expect that about $(0.05)(40) = 2$ of the intervals so constructed would fail to contain the population mean μ, and in this simulation two of the intervals, shown in red, do.

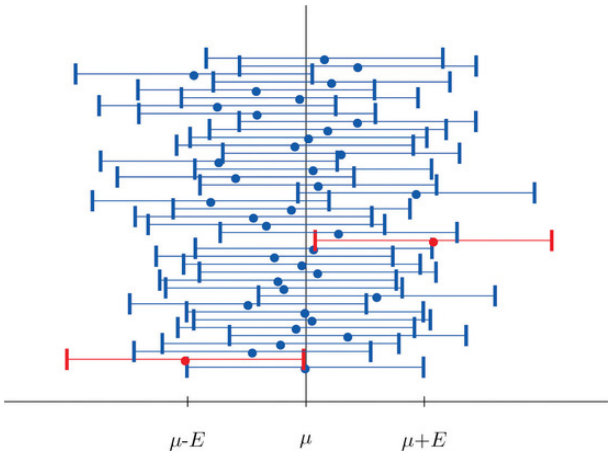


Figure 7.2.1.2: Computer Simulation of 40 95% Confidence Intervals for a Mean

It is standard practice to identify the level of confidence in terms of the area α in the two tails of the distribution of $\overline{X}$ when the middle part specified by the level of confidence is taken out. This is shown in Figure 7.2.1.3, drawn for the general situation, and in Figure 7.2.1.4, drawn for 95% confidence.

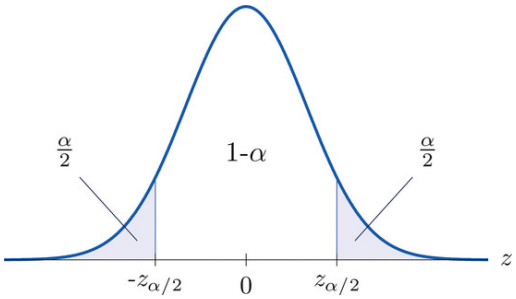


Figure 7.2.1.3: For $100(1-\alpha)\%$ confidence the area in each tail is $\alpha/2$.

Remember from Section 5.4 that the z-value that cuts off a right tail of area c is denoted z_c. Thus the number 1.960 in the example is $z_{.025}$, which is $z_{\frac{\alpha}{2}}$ for $\alpha = 1 - 0.95 = 0.05$.

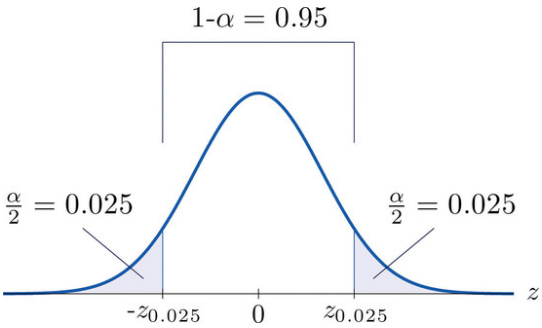


Figure 7.2.1.4: For 95% confidence the area in each tail is $\alpha/2 = 0.025$.

For 95% confidence the area in each tail is $\alpha/2 = 0.025$.

The level of confidence can be any number between 0 and 100%, but the most common values are probably 90% ($\alpha = 0.10$), 95% ($\alpha = 0.05$), and 99% ($\alpha = 0.01$).

Thus in general for a $100(1-\alpha)\%$ confidence interval, $E = z_{\alpha/2}(\sigma/\sqrt{n})$, so the formula for the confidence interval is $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$. While sometimes the population standard deviation σ is known, typically it is not. If not, for $n \geq 30$ it is generally safe to approximate σ by the sample standard deviation s.

Large Sample $100(1-\alpha)\%$ Confidence Interval for a Population Mean

- If σ is known:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- If σ is unknown:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

A sample is considered large when $n \geq 30$.

As mentioned earlier, the number

$$E = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

or

$$E = z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

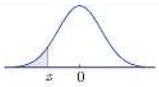
is called the *margin of error of the estimate*.

Example 7.2.1.1

Find the number $z_{\alpha/2}$ needed in construction of a confidence interval:

1. when the level of confidence is 90%;
2. when the level of confidence is 99%.

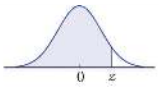
using the tables in Figure 7.2.1.5 below.



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

https://stats.libretexts.org/@go/page/48565



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8304	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure **7.2.1.5**: Cumulative Normal Probability

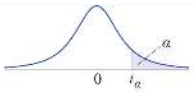
Solution:

1. For confidence level 90%, $\alpha = 1 - 0.90 = 0.10$, so $z_{\alpha/2} = z_{0.05}$. Since the area under the standard normal curve to the right of $z_{0.05}$ is 0.05, the area to the left of $z_{0.05}$ is 0.95. We search for the area 0.9500 in Figure **7.2.1.5**. The closest entries in the table are 0.9495 and 0.9505, corresponding to z-values 1.64 and 1.65. Since 0.95 is halfway between 0.9495 and 0.9505 we use the average 1.645 of the z-values for $z_{0.05}$.
2. For confidence level 99%, $\alpha = 1 - 0.99 = 0.01$, so $z_{\alpha/2} = z_{0.005}$. Since the area under the standard normal curve to the right of $z_{0.005}$ is 0.005, the area to the left of $z_{0.005}$ is 0.9950. We search for the area 0.9950 in Figure **7.2.1.5**. The closest entries in the table are 0.9949 and 0.9951, corresponding to z-values 2.57 and 2.58. Since 0.995 is halfway between 0.9949 and 0.9951 we use the average 2.575 of the z-values for $z_{0.005}$.

Example 7.2.1.2

Use Figure **7.2.1.6** below to find the number $z_{\alpha/2}$ needed in construction of a confidence interval:

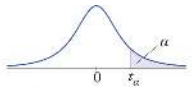
1. when the level of confidence is 90%;
2. when the level of confidence is 99%.



Critical Values of t

df	$t_{0.200}$	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.0025}$	$t_{0.001}$	$t_{0.0005}$
1	1.376	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.061	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.978	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.920	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.876	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.873	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.870	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.868	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.866	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.865	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.863	1.333	1.740	2.110	2.576	2.898	3.222	3.646	3.965
18	0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.859	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.858	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.858	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.857	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.856	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.856	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.855	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.855	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.854	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.854	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.853	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.853	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.853	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.852	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.852	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.852	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.851	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.851	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39	0.851	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.851	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
41	0.851	1.303	1.683	2.020	2.421	2.701	2.967	3.301	3.544
42	0.851	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
43	0.851	1.302	1.681	2.017	2.416	2.695	2.959	3.291	3.532
44	8.850	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
45	0.850	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
46	0.850	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
47	0.849	1.300	1.678	2.012	2.408	2.685	2.946	3.273	3.510
48	0.849	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
49	0.849	1.299	1.677	2.010	2.405	2.680	2.940	3.265	3.500
50	0.849	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496

https://stats.libretexts.org/@go/page/48565



Critical Values of *t*

df	$t_{0.200}$	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.0025}$	$t_{0.001}$	$t_{0.0005}$
51	0.849	1.298	1.675	2.008	2.402	2.676	2.934	3.258	3.492
52	0.849	1.298	1.675	2.007	2.400	2.674	2.932	3.255	3.488
53	0.849	1.298	1.674	2.006	2.399	2.672	2.929	3.251	3.484
54	0.848	1.297	1.674	2.005	2.397	2.670	2.927	3.248	3.480
55	0.848	1.297	1.673	2.004	2.396	2.668	2.925	3.245	3.476
56	0.848	1.297	1.673	2.003	2.395	2.667	2.923	3.242	3.473
57	0.848	1.297	1.672	2.002	2.394	2.665	2.920	3.239	3.470
58	0.848	1.296	1.672	2.002	2.392	2.663	2.918	3.237	3.466
59	0.848	1.296	1.671	2.001	2.391	2.662	2.916	3.234	3.463
60	0.848	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
61	0.848	1.296	1.670	2.000	2.389	2.659	2.913	3.229	3.457
62	0.848	1.295	1.670	1.999	2.388	2.657	2.911	3.227	3.454
63	0.847	1.295	1.669	1.998	2.387	2.656	2.909	3.225	3.452
64	0.847	1.295	1.669	1.998	2.386	2.655	2.908	3.223	3.449
65	0.847	1.295	1.669	1.997	2.385	2.654	2.906	3.220	3.447
66	0.847	1.295	1.668	1.997	2.384	2.652	2.904	3.218	3.444
67	0.847	1.294	1.668	1.996	2.383	2.651	2.903	3.216	3.442
68	0.847	1.294	1.668	1.995	2.382	2.650	2.902	3.214	3.439
69	0.847	1.294	1.667	1.995	2.382	2.649	2.900	3.213	3.437
70	0.847	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
71	0.847	1.294	1.667	1.994	2.380	2.647	2.897	3.209	3.433
72	0.847	1.293	1.666	1.993	2.379	2.646	2.896	3.207	3.431
73	0.847	1.293	1.666	1.993	2.379	2.645	2.895	3.206	3.429
74	0.847	1.293	1.666	1.993	2.378	2.644	2.894	3.204	3.427
75	0.846	1.293	1.665	1.992	2.377	2.643	2.892	3.202	3.425
76	0.846	1.293	1.665	1.992	2.376	2.642	2.891	3.201	3.423
77	0.846	1.293	1.665	1.991	2.376	2.641	2.890	3.199	3.421
78	0.846	1.292	1.665	1.991	2.375	2.640	2.889	3.198	3.420
79	0.846	1.292	1.664	1.990	2.374	2.640	2.888	3.197	3.418
80	0.846	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
81	0.846	1.292	1.664	1.990	2.373	2.638	2.886	3.194	3.415
82	0.846	1.292	1.664	1.989	2.373	2.637	2.885	3.193	3.413
83	0.846	1.292	1.663	1.989	2.372	2.636	2.884	3.191	3.412
84	0.846	1.292	1.663	1.989	2.372	2.636	2.883	3.190	3.410
85	0.846	1.292	1.663	1.988	2.371	2.635	2.882	3.189	3.409
86	0.846	1.291	1.663	1.988	2.370	2.634	2.881	3.188	3.407
87	0.846	1.291	1.663	1.988	2.370	2.634	2.880	3.187	3.406
88	0.846	1.291	1.662	1.987	2.369	2.633	2.880	3.185	3.405
89	0.846	1.291	1.662	1.987	2.369	2.632	2.879	3.184	3.403
90	0.846	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
91	0.846	1.291	1.662	1.986	2.368	2.631	2.877	3.182	3.401
92	0.846	1.291	1.662	1.986	2.368	2.630	2.876	3.181	3.399
93	0.846	1.291	1.661	1.986	2.367	2.630	2.876	3.180	3.398
94	0.846	1.291	1.661	1.986	2.367	2.629	2.875	3.179	3.397
95	0.845	1.291	1.661	1.985	2.366	2.629	2.874	3.178	3.396
96	0.845	1.290	1.661	1.985	2.366	2.628	2.873	3.177	3.395
97	0.845	1.290	1.661	1.985	2.365	2.627	2.873	3.176	3.394
98	0.845	1.290	1.661	1.984	2.365	2.627	2.872	3.175	3.393
99	0.845	1.290	1.660	1.984	2.365	2.626	2.871	3.175	3.392
100	0.845	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
∞ [*z*]	0.842	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Figure **7.2.1.6**: Critical Values of t

Solution:

1. In the next section we will learn about a continuous random variable that has a probability distribution called the Student t-distribution. Figure **7.2.1.6** gives the value t_c that cuts off a right tail of area c for different values of c. The last line of that table, the one whose heading is the symbol ∞ for infinity and $[z]$, gives the corresponding z-value z_c that cuts off a right tail of the same area c. In particular, $z_{0.05}$ is the number in that row and in the column with the heading $t_{0.05}$. We read off directly that $z_{0.05} = 1.645$.
2. In Figure **7.2.1.6** $z_{0.005}$ is the number in the last row and in the column headed $t_{0.005}$, namely **2.576**.

Figure **7.2.1.6** can be used to find z_c only for those values of c for which there is a column with the heading t_c appearing in the table; otherwise we must use Figure **7.2.1.5** in reverse. But when it can be done it is both faster and more accurate to use the last line of Figure **7.2.1.6** to find z_c than it is to do so using Figure **7.2.1.5** in reverse.

Example 7.2.1.3

A sample of size 49 has sample mean 35 and sample standard deviation 14. Construct a 98% confidence interval for the population mean using this information. Interpret its meaning.

Solution:

For confidence level 98%, $\alpha = 1 - 0.98 = 0.02$, so $z_{\alpha/2} = z_{0.01}$. From Figure $7.2.1.6$ we read directly that $z_{0.01} = 2.326$.Thus

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 35 \pm 2.326\left(\frac{14}{\sqrt{49}}\right) = 35 \pm 4.652 \approx 35 \pm 4.7$$

We are 98% confident that the population mean μ lies in the interval $[30.3, 39.7]$, in the sense that in repeated sampling 98% of all intervals constructed from the sample data in this manner will contain μ.

Example 7.2.1.4

A random sample of 120 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university.

Solution:

For confidence level 90%, $\alpha = 1 - 0.90 = 0.10$, so $z_{\alpha/2} = z_{0.05}$. From Figure $7.2.1.6$ we read directly that $z_{0.05} = 1.645$. Since $n = 120$, $\bar{x} = 2.71$, and $s = 0.51$,

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 2.71 \pm 1.645\left(\frac{0.51}{\sqrt{120}}\right) = 2.71 \pm 0.0766$$

One may be 90% confident that the true average GPA of all students at the university is contained in the interval $(2.71 - 0.08, 2.71 + 0.08) = (2.63, 2.79)$.

- A confidence interval for a population mean is an estimate of the population mean together with an indication of reliability.
- There are different formulas for a confidence interval based on the sample size and whether or not the population standard deviation is known.
- The confidence intervals are constructed entirely from the sample data (or sample data and the population standard deviation, when it is known).

This page titled 7.2.1: Large Sample Estimation of a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **7.1: Large Sample Estimation of a Population Mean** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

7.2.2: Small Sample Estimation of a Population Mean

Learning Objectives

1. To become familiar with Student's t-distribution.
2. To understand how to apply additional formulas for a confidence interval for a population mean.

The confidence interval formulas in the previous section are based on the Central Limit Theorem, the statement that for large samples $\overline{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. When the population mean μ is estimated with a small sample ($n < 30$), the Central Limit Theorem does not apply. In order to proceed we assume that the numerical population from which the sample is taken has a normal distribution to begin with. If this condition is satisfied then when the population standard deviation σ is known the old formula $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ can still be used to construct a $100(1-\alpha)\%$ confidence interval for μ.

If the population standard deviation is unknown and the sample size n is small then when we substitute the sample standard deviation s for σ the normal approximation is no longer valid. The solution is to use a different distribution, called Student's t-distribution with $n-1$ degrees of freedom. Student's t-distribution is very much like the standard normal distribution in that it is centered at 0 and has the same qualitative bell shape, but it has heavier tails than the standard normal distribution does, as indicated by Figure 7.2.2.1, in which the curve (in brown) that meets the dashed vertical line at the lowest point is the t-distribution with two degrees of freedom, the next curve (in blue) is the t-distribution with five degrees of freedom, and the thin curve (in red) is the standard normal distribution. As also indicated by the figure, as the sample size n increases, Student's t-distribution ever more closely resembles the standard normal distribution. Although there is a different t-distribution for every value of n, once the sample size is 30 or more it is typically acceptable to use the standard normal distribution instead, as we will always do in this text.

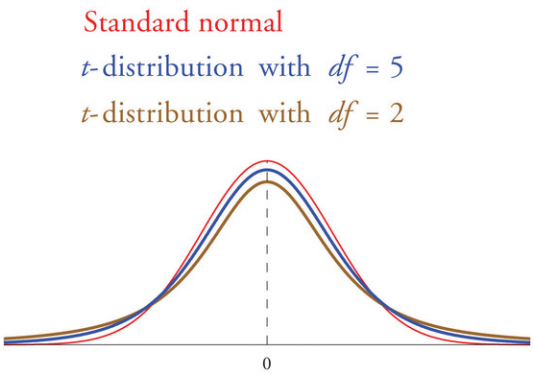


Figure 7.2.2.1: Student's t-*Distribution*

Just as the symbol z_c stands for the value that cuts off a right tail of area c in the standard normal distribution, so the symbol t_c stands for the value that cuts off a right tail of area c in the standard normal distribution. This gives us the following confidence interval formulas.

Small Sample $100(1-\alpha)\%$ Confidence Interval for a Population Mean

If σ is known:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

If σ is unknown:

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \qquad (7.2.2.1)$$

with the degrees of freedom $df = n-1$.

The population must be normally distributed and a sample is considered small when $n < 30$.

To use the new formula we use the line in Figure 7.1.6 that corresponds to the relevant sample size.

Example 7.2.2.1

A sample of size 15 drawn from a normally distributed population has sample mean 35 and sample standard deviation 14. Construct a 95% confidence interval for the population mean, and interpret its meaning.

Solution

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is Equation 7.2.2.1.

Confidence level 95% means that

$$\alpha = 1 - 0.95 = 0.05$$

so $\alpha/2 = 0.025$. Since the sample size is $n = 15$, there are $n - 1 = 14$ degrees of freedom. By Figure 7.1.6 $t_{0.025} = 2.145$. Thus

$$\overline{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \tag{7.2.2.2}$$

$$= 35 \pm 2.145\left(\frac{14}{\sqrt{15}}\right) \tag{7.2.2.3}$$

$$= 35 \pm 7.8 \tag{7.2.2.4}$$

One may be 95% confident that the true value of μ is contained in the interval

$$(35 - 7.8, 35 + 7.8) = (27.2, 42.8).$$

Example 7.2.2.2

A random sample of 12 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university. Assume that the numerical population of GPAs from which the sample is taken has a normal distribution.

Solution

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is Equation 7.2.2.1

Confidence level 90% means that

$$\alpha = 1 - 0.90 = 0.10$$

so $\alpha/2 = 0.05$. Since the sample size is $n = 12$, there are $n - 1 = 11$ degrees of freedom. By Figure 7.1.6 $t_{0.05} = 1.796$. Thus

$$\overline{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \tag{7.2.2.5}$$

$$= 2.71 \pm 1.796\left(\frac{0.51}{\sqrt{12}}\right) \tag{7.2.2.6}$$

$$= 2.71 \pm 0.26 \tag{7.2.2.7}$$

One may be 90% confident that the true average GPA of all students at the university is contained in the interval

$$(2.71 - 0.26, 2.71 + 0.26) = (2.45, 2.97).$$

Compare "Example 4" in Section 7.1 and "Example 6" in Section 7.1. The summary statistics in the two samples are the same, but the 90% confidence interval for the average GPA of all students at the university in "Example 4" in Section 7.1, $(2.63, 2.79)$, is shorter than the 90% confidence interval $(2.45, 2.97)$, in "Example 6" in Section 7.1. This is partly because in "Example 4" in Section 7.1 the sample size is larger; there is more information pertaining to the true value of μ in the large data set than in the small one.

Key Takeaway

- In selecting the correct formula for construction of a confidence interval for a population mean ask two questions: is the population standard deviation σ known or unknown, and is the sample large or small?
- We can construct confidence intervals with small samples only if the population is normal.

This page titled 7.2.2: Small Sample Estimation of a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **7.2: Small Sample Estimation of a Population Mean** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

7.3: Sample Size Considerations

Learning Objectives

- To learn how to apply formulas for estimating the size sample that will be needed in order to construct a confidence interval for a population mean or proportion that meets given criteria.

Sampling is typically done with a set of clear objectives in mind. For example, an economist might wish to estimate the mean yearly income of workers in a particular industry at 90% confidence and to within $\$500$. Since sampling costs time, effort, and money, it would be useful to be able to estimate the smallest size sample that is likely to meet these criteria.

Estimating μ

The confidence interval formulas for estimating a population mean μ have the form $\overline{x} \pm E$. When the population standard deviation σ is known,

$$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

The number $z_{\alpha/2}$ is determined by the desired level of confidence. To say that we wish to estimate the mean to within a certain number of units means that we want the margin of error E to be no larger than that number. Thus we obtain the minimum sample size needed by solving the displayed equation for n.

Minimum Sample Size for Estimating a Population Mean

The estimated minimum sample size n needed to estimate a population mean μ to within E units at $100(1-\alpha)\%$ confidence is

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \text{ (rounded up)} \tag{7.3.1}$$

To apply Equation 7.3.1, we must have prior knowledge of the population in order to have an estimate of its standard deviation σ. In all the examples and exercises the population standard deviation will be given.

Example 7.3.1

Find the minimum sample size necessary to construct a 99% confidence interval for μ with a margin of error $E = 0.2$. Assume that the population standard deviation is $\sigma = 1.3$.

Solution

Confidence level 99% means that $\alpha = 1 - 0.99 = 0.01$ so $\alpha/2 = 0.005$. From the last line of Figure 7.1.6 we obtain $z_{0.005} = 2.576$. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(2.576)^2(1.3)^2}{(0.2)^2} = 280.361536$$

which we round up to 281, since it is impossible to take a fractional observation.

Example 7.3.2

An economist wishes to estimate, with a 95% confidence interval, the yearly income of welders with at least five years experience to within $\$1,000$. He estimates that the range of incomes is no more than $\$24,000$, so using the Empirical Rule he estimates the population standard deviation to be about one-sixth as much, or about $\$4,000$. Find the estimated minimum sample size required.

Solution

Confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$ so $\alpha/2 = 0.025$. From the last line of Figure 7.1.6 we obtain $z_{0.025} = 1.960$.

To say that the estimate is to be "to within $\$1,000$" means that $E = 1000$. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.960)^2(4000)^2}{(1000)^2} = 61.4656$$

which we round up to 62.

Estimating p

The confidence interval formula for estimating a population proportion p is $\hat{p} \pm E$, where

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The number $z_{\alpha/2}$ is determined by the desired level of confidence. To say that we wish to estimate the population proportion to within a certain number of percentage points means that we want the margin of error E to be no larger than that number (expressed as a proportion). Thus we obtain the minimum sample size needed by solving the displayed equation for n.

Minimum Sample Size for Estimating a Population Proportion

The estimated minimum sample size n needed to estimate a population proportion p to within E at $100(1-\alpha)\%$ confidence is

$$n = \frac{(z_{\alpha/2})^2\hat{p}(1-\hat{p})}{E^2}(\text{rounded up})$$

There is a dilemma here: the formula for estimating how large a sample to take contains the number $\hat{p}$, which we know only after we have taken the sample. There are two ways out of this dilemma. Typically the researcher will have some idea as to the value of the population proportion p, hence of what the sample proportion $\hat{p}$ is likely to be. For example, if last month 37% of all voters thought that state taxes are too high, then it is likely that the proportion with that opinion this month will not be dramatically different, and we would use the value 0.37 for $\hat{p}$ in the formula.

The second approach to resolving the dilemma is simply to replace $\hat{p}$ in the formula by 0.5. This is because if $\hat{p}$ is large then $1-\hat{p}$ is small, and vice versa, which limits their product to a maximum value of 0.25, which occurs when $\hat{p} = 0.5$. This is called the **most conservative estimate**, since it gives the largest possible estimate of n.

Example 7.3.3

Find the necessary minimum sample size to construct a 98% confidence interval for p with a margin of error $E = 0.05$,

a. assuming that no prior knowledge about p is available; and
b. assuming that prior studies suggest that p is about 0.1.

Solution

Confidence level 98% means that $\alpha = 1 - 0.98 = 0.02$ so $\alpha/2 = 0.01$. From the last line of Figure 7.1.6 we obtain $z_{0.01} = 2.326$.

a. Since there is no prior knowledge of p we make the most conservative estimate that $\hat{p} = 0.5$. Then

$$n = \frac{(z_{\alpha/2})^2\hat{p}(1-\hat{p})}{E^2} = \frac{(2.326)^2(0.5)(1-0.5)}{0.05^2} = 541.0276$$

which we round up to 542.

b. Since $p \approx 0.1$ we estimate $\hat{p}$ by 0.1, and obtain

$$n = \frac{(z_{\alpha/2})^2\hat{p}(1-\hat{p})}{E^2} = \frac{(2.326)^2(0.1)(1-0.1)}{0.05^2} = 194.769936$$

which we round up to 195.

Example 7.3.4

A dermatologist wishes to estimate the proportion of young adults who apply sunscreen regularly before going out in the sun in the summer. Find the minimum sample size required to estimate the proportion to within three percentage points, at 90% confidence.

Solution

Confidence level 90% means that $\alpha = 1 - 0.90 = 0.10$ so $\alpha/2 = 0.05$. From the last line of Figure 7.1.6 we obtain $z_{0.05} = 1.645$.

Since there is no prior knowledge of p we make the most conservative estimate that $\hat{p} = 0.5$. To estimate "to within three percentage points" means that $E = 0.03$. Then

$$n = \frac{(z_{\alpha/2})^2 \hat{p}(1-\hat{p})}{E^2} = \frac{(1.645)^2(0.5)(1-0.5)}{0.03^2} = 751.6736111$$

which we round up to 752.

Key Takeaway

- If the population standard deviation σ is known or can be estimated, then the minimum sample size needed to obtain a confidence interval for the population mean with a given maximum error of the estimate and a given level of confidence can be estimated.
- The minimum sample size needed to obtain a confidence interval for a population proportion with a given maximum error of the estimate and a given level of confidence can always be estimated. If there is prior knowledge of the population proportion p then the estimate can be sharpened.

This page titled 7.3: Sample Size Considerations is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **7.4: Sample Size Considerations** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

7.E: Estimation (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

7.1: Large Sample Estimation of a Population Mean

Basic

1. A random sample is drawn from a population of known standard deviation 11.3. Construct a 90% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 36,\ \bar{x} = 105.2,\ s = 11.2$
   b. $n = 100,\ \bar{x} = 105.2,\ s = 11.2$
2. A random sample is drawn from a population of known standard deviation 22.1. Construct a 95% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 121,\ \bar{x} = 82.4,\ s = 21.9$
   b. $n = 81,\ \bar{x} = 82.4,\ s = 21.9$
3. A random sample is drawn from a population of unknown standard deviation. Construct a 99% confidence interval for the population mean based on the information given.
   a. $n = 49,\ \bar{x} = 17.1,\ s = 2.1$
   b. $n = 169,\ \bar{x} = 17.1,\ s = 2.1$
4. A random sample is drawn from a population of unknown standard deviation. Construct a 98% confidence interval for the population mean based on the information given.
   a. $n = 225,\ \bar{x} = 92.0,\ s = 8.4$
   b. $n = 64,\ \bar{x} = 92.0,\ s = 8.4$
5. A random sample of size 144 is drawn from a population whose distribution, mean, and standard deviation are all unknown. The summary statistics are $\bar{x} = 58.2$ and $s = 2.6$.
   a. Construct an 80% confidence interval for the population mean μ.
   b. Construct a 90% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.
6. A random sample of size 256 is drawn from a population whose distribution, mean, and standard deviation are all unknown. The summary statistics are $\bar{x} = 1011$ and $s = 34$.
   a. Construct a 90% confidence interval for the population mean μ.
   b. Construct a 99% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.

Applications

7. A government agency was charged by the legislature with estimating the length of time it takes citizens to fill out various forms. Two hundred randomly selected adults were timed as they filled out a particular form. The times required had mean 12.8 minutes with standard deviation 1.7 minutes. Construct a 90% confidence interval for the mean time taken for all adults to fill out this form.
8. Four hundred randomly selected working adults in a certain state, including those who worked at home, were asked the distance from their home to their workplace. The average distance was 8.84 miles with standard deviation 2.70 miles. Construct a 99% confidence interval for the mean distance from home to work for all residents of this state.
9. On every passenger vehicle that it tests an automotive magazine measures, at true speed 55 mph, the difference between the true speed of the vehicle and the speed indicated by the speedometer. For 36 vehicles tested the mean difference was -1.2 mph with standard deviation 0.2 mph. Construct a 90% confidence interval for the mean difference between true speed and indicated speed for all vehicles.
10. A corporation monitors time spent by office workers browsing the web on their computers instead of working. In a sample of computer records of 50 workers, the average amount of time spent browsing in an eight-hour work day was 27.8 minutes with

standard deviation 8.2 minutes. Construct a 99.5% confidence interval for the mean time spent by all office workers in browsing the web in an eight-hour day.

11. A sample of 250 workers aged 16 and older produced an average length of time with the current employer ("job tenure") of 4.4 years with standard deviation 3.8 years. Construct a 99.9% confidence interval for the mean job tenure of all workers aged 16 or older.
12. The amount of a particular biochemical substance related to bone breakdown was measured in 30 healthy women. The sample mean and standard deviation were 3.3 nanograms per milliliter (ng/mL) and 1.4 ng/mL. Construct an 80% confidence interval for the mean level of this substance in all healthy women.
13. A corporation that owns apartment complexes wishes to estimate the average length of time residents remain in the same apartment before moving out. A sample of 150 rental contracts gave a mean length of occupancy of 3.7 years with standard deviation 1.2 years. Construct a 95% confidence interval for the mean length of occupancy of apartments owned by this corporation.
14. The designer of a garbage truck that lifts roll-out containers must estimate the mean weight the truck will lift at each collection point. A random sample of 325 containers of garbage on current collection routes yielded $\bar{x} = 75.3lb,\ s = 12.8lb$. Construct a 99.8% confidence interval for the mean weight the trucks must lift each time.
15. In order to estimate the mean amount of damage sustained by vehicles when a deer is struck, an insurance company examined the records of 50 such occurrences, and obtained a sample mean of $\$2,785$ with sample standard deviation $\$221$. Construct a 95% confidence interval for the mean amount of damage in all such accidents.
16. In order to estimate the mean FICO credit score of its members, a credit union samples the scores of 95 members, and obtains a sample mean of 738.2 with sample standard deviation 64.2. Construct a 99% confidence interval for the mean FICO score of all of its members.

Additional Exercises

17. For all settings a packing machine delivers a precise amount of liquid; the amount dispensed always has standard deviation 0.07 ounce. To calibrate the machine its setting is fixed and it is operated 50 times. The mean amount delivered is 6.02 ounces with sample standard deviation 0.04 ounce. Construct a 99.5% confidence interval for the mean amount delivered at this setting. Hint: Not all the information provided is needed.
18. A power wrench used on an assembly line applies a precise, preset amount of torque; the torque applied has standard deviation 0.73 foot-pound at every torque setting. To check that the wrench is operating within specifications it is used to tighten 100 fasteners. The mean torque applied is 36.95 foot-pounds with sample standard deviation 0.62 foot-pound. Construct a 99.9% confidence interval for the mean amount of torque applied by the wrench at this setting. Hint: Not all the information provided is needed.
19. The number of trips to a grocery store per week was recorded for a randomly selected collection of households, with the results shown in the table.

$$\begin{array}{cccccccccc} 2 & 2 & 2 & 1 & 4 & 2 & 3 & 2 & 5 & 4 \\ 2 & 3 & 5 & 0 & 3 & 2 & 3 & 1 & 4 & 3 \\ 3 & 2 & 1 & 6 & 2 & 3 & 3 & 2 & 4 & 4 \end{array} \tag{7.E.1}$$

Construct a 95% confidence interval for the average number of trips to a grocery store per week of all households.

20. For each of 40 high school students in one county the number of days absent from school in the previous year were counted, with the results shown in the frequency table.

x	0	1	2	3	4	5
f	24	7	5	2	1	1

(7.E.2)

Construct a 90% confidence interval for the average number of days absent from school of all students in the county.

21. A town council commissioned a random sample of 85 households to estimate the number of four-wheel vehicles per household in the town. The results are shown in the following frequency table.

x	0	1	2	3	4	5
f	1	16	28	22	12	6

(7.E.3)

Construct a 98% confidence interval for the average number of four-wheel vehicles per household in the town.

22. The number of hours per day that a television set was operating was recorded for a randomly selected collection of households, with the results shown in the table.

3.7	4.2	1.5	3.6	5.9
4.7	8.2	3.9	2.5	4.4
2.1	3.6	1.1	7.3	4.2
3.0	3.8	2.2	4.2	3.8
4.3	2.1	2.4	6.0	3.7
2.5	1.3	2.8	3.0	5.6

(7.E.4)

Construct a 99.8% confidence interval for the mean number of hours that a television set is in operation in all households.

Large Data Set Exercises

Large Data Set missing from the original

23. Large **Data Set 1** records the SAT scores of 1,000 students. Regarding it as a random sample of all high school students, use it to construct a 99% confidence interval for the mean SAT score of all students.
24. Large **Data Set 1** records the GPAs of 1,000 college students. Regarding it as a random sample of all college students, use it to construct a 95% confidence interval for the mean GPA of all students.
25. Large **Data Set 1** lists the SAT scores of 1,000 students.
    a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.
    b. Regard the first 36 students as a random sample and use it to construct a 99% confidence for the mean μ of all 1,000 SAT scores. Does it actually capture the mean μ?
26. Large **Data Set 1** lists the GPAs of 1,000 students.
    a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.
    b. Regard the first 36 students as a random sample and use it to construct a 95% confidence for the mean μ of all 1,000 GPAs. Does it actually capture the mean μ?

Answers

1. a. 105.2 ± 3.10
   b. 105.2 ± 1.86
2.
3. a. 17.1 ± 0.77
   b. 17.1 ± 0.42
4.
5. a. 58.2 ± 0.28
   b. 58.2 ± 0.36
   c. Asking for greater confidence requires a longer interval.
6.
7. 12.8 ± 0.20
8.
9. -1.2 ± 0.05
10.
11. 4.4 ± 0.79
12.
13. 3.7 ± 0.19
14.
15. 2785 ± 61
16.

17. 6.02 ± 0.03
18.
19. 2.8 ± 0.48
20.
21. 2.54 ± 0.30
22.
23. $(1511.43, 1546.05)$
24.
25. a. $\mu = 1528.74$
    b. $(1428.22, 1602.89)$

7.2: Small Sample Estimation of a Population Mean

Basic

1. A random sample is drawn from a normally distributed population of known standard deviation 5. Construct a 99.8% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 16,\ \bar{x} = 98,\ s = 5.6$
   b. $n = 9,\ \bar{x} = 98,\ s = 5.6$
2. A random sample is drawn from a normally distributed population of known standard deviation 10.7. Construct a 95% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 25,\ \bar{x} = 103.3,\ s = 11.0$
   b. $n = 4,\ \bar{x} = 103.3,\ s = 11.0$
3. A random sample is drawn from a normally distributed population of unknown standard deviation. Construct a 99% confidence interval for the population mean based on the information given.
   a. $n = 18,\ \bar{x} = 386,\ s = 24$
   b. $n = 7,\ \bar{x} = 386,\ s = 24$
4. A random sample is drawn from a normally distributed population of unknown standard deviation. Construct a 98% confidence interval for the population mean based on the information given.
   a. $n = 8,\ \bar{x} = 58.3,\ s = 4.1$
   b. $n = 27,\ \bar{x} = 58.3,\ s = 4.1$
5. A random sample of size 14 is drawn from a normal population. The summary statistics are $\bar{x} = 933,\ and\ s = 18$.
   a. Construct an 80% confidence interval for the population mean μ.
   b. Construct a 90% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.
6. A random sample of size 28 is drawn from a normal population. The summary statistics are $\bar{x} = 68.6,\ and\ s = 1.28$.
   a. Construct a 95% confidence interval for the population mean μ.
   b. Construct a 99.5% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.

Application Exercises

7. City planners wish to estimate the mean lifetime of the most commonly planted trees in urban settings. A sample of 16 recently felled trees yielded mean age 32.7 years with standard deviation 3.1 years. Assuming the lifetimes of all such trees are normally distributed, construct a 99.8% confidence interval for the mean lifetime of all such trees.
8. To estimate the number of calories in a cup of diced chicken breast meat, the number of calories in a sample of four separate cups of meat is measured. The sample mean is 211.8 calories with sample standard deviation 0.9 calorie. Assuming the caloric content of all such chicken meat is normally distributed, construct a 95% confidence interval for the mean number of calories in one cup of meat.
9. A college athletic program wishes to estimate the average increase in the total weight an athlete can lift in three different lifts after following a particular training program for six weeks. Twenty-five randomly selected athletes when placed on the program

exhibited a mean gain of 47.3 lb with standard deviation 6.4 lb. Construct a 90% confidence interval for the mean increase in lifting capacity all athletes would experience if placed on the training program. Assume increases among all athletes are normally distributed.

10. To test a new tread design with respect to stopping distance, a tire manufacturer manufactures a set of prototype tires and measures the stopping distance from 70 mph on a standard test car. A sample of 25 stopping distances yielded a sample mean 173 feet with sample standard deviation 8 feet. Construct a 98% confidence interval for the mean stopping distance for these tires. Assume a normal distribution of stopping distances.
11. A manufacturer of chokes for shotguns tests a choke by shooting 15 patterns at targets 40 yards away with a specified load of shot. The mean number of shot in a 30-inch circle is 53.5 with standard deviation 1.6. Construct an 80% confidence interval for the mean number of shot in a 30-inch circle at 40 yards for this choke with the specified load. Assume a normal distribution of the number of shot in a 30-inch circle at 40 yards for this choke.
12. In order to estimate the speaking vocabulary of three-year-old children in a particular socioeconomic class, a sociologist studies the speech of four children. The mean and standard deviation of the sample are $\bar{x} = 1120$ and $s = 215$ words. Assuming that speaking vocabularies are normally distributed, construct an 80% confidence interval for the mean speaking vocabulary of all three-year-old children in this socioeconomic group.
13. A thread manufacturer tests a sample of eight lengths of a certain type of thread made of blended materials and obtains a mean tensile strength of 8.2 lb with standard deviation 0.06 lb. Assuming tensile strengths are normally distributed, construct a 90% confidence interval for the mean tensile strength of this thread.
14. An airline wishes to estimate the weight of the paint on a fully painted aircraft of the type it flies. In a sample of four repaintings the average weight of the paint applied was 239 pounds, with sample standard deviation 8 pounds. Assuming that weights of paint on aircraft are normally distributed, construct a 99.8% confidence interval for the mean weight of paint on all such aircraft.
15. In a study of dummy foal syndrome, the average time between birth and onset of noticeable symptoms in a sample of six foals was 18.6 hours, with standard deviation 1.7 hours. Assuming that the time to onset of symptoms in all foals is normally distributed, construct a 90% confidence interval for the mean time between birth and onset of noticeable symptoms.
16. A sample of 26 women's size 6 dresses had mean waist measurement 25.25 inches with sample standard deviation 0.375 inch. Construct a 95% confidence interval for the mean waist measurement of all size 6 women's dresses. Assume waist measurements are normally distributed.

Additional Exercises

17. Botanists studying attrition among saplings in new growth areas of forests diligently counted stems in six plots in five-year-old new growth areas, obtaining the following counts of stems per acre:

$$\begin{array}{ccc} 9,432 & 11,026 & 10,539 \\ 8,773 & 9,868 & 10,247 \end{array} \tag{7.E.5}$$

Construct an 80% confidence interval for the mean number of stems per acre in all five-year-old new growth areas of forests. Assume that the number of stems per acre is normally distributed.

18. Nutritionists are investigating the efficacy of a diet plan designed to increase the caloric intake of elderly people. The increase in daily caloric intake in 12 individuals who are put on the plan is (a minus sign signifies that calories consumed went down):

$$\begin{array}{cccccc} 121 & 284 & -94 & 295 & 183 & 312 \\ 188 & -102 & 259 & 226 & 152 & 167 \end{array} \tag{7.E.6}$$

Construct a 99.8% confidence interval for the mean increase in caloric intake for all people who are put on this diet. Assume that population of differences in intake is normally distributed.

19. A machine for making precision cuts in dimension lumber produces studs with lengths that vary with standard deviation 0.003 inch. Five trial cuts are made to check the machine's calibration. The mean length of the studs produced is 104.998 inches with sample standard deviation 0.004 inch. Construct a 99.5% confidence interval for the mean lengths of all studs cut by this machine. Assume lengths are normally distributed. Hint: Not all the numbers given in the problem are used.

20. The variation in time for a baked good to go through a conveyor oven at a large scale bakery has standard deviation 0.017 minute at every time setting. To check the bake time of the oven periodically four batches of goods are carefully timed. The recent check gave a mean of 27.2 minutes with sample standard deviation 0.012 minute. Construct a 99.8% confidence interval for the mean bake time of all batches baked in this oven. Assume bake times are normally distributed. Hint: Not all the numbers given in the problem are used.
21. Wildlife researchers tranquilized and weighed three adult male polar bears. The data (in pounds) are: $926, 742, 1109$. Assume the weights of all bears are normally distributed.
    a. Construct an 80% confidence interval for the mean weight of all adult male polar bears using these data.
    b. Convert the three weights in pounds to weights in kilograms using the conversion $1\ lb = 0.453\ kg$ (so the first datum changes to $(926)(0.453) = 419$). Use the converted data to construct an 80% confidence interval for the mean weight of all adult male polar bears expressed in kilograms.
    c. Convert your answer in part (a) into kilograms directly and compare it to your answer in (b). This illustrates that if you construct a confidence interval in one system of units you can convert it directly into another system of units without having to convert all the data to the new units.
22. Wildlife researchers trapped and measured six adult male collared lemmings. The data (in millimeters) are: $104, 99, 112, 115, 96, 109$. Assume the lengths of all lemmings are normally distributed.
    a. Construct a 90% confidence interval for the mean length of all adult male collared lemmings using these data.
    b. Convert the six lengths in millimeters to lengths in inches using the conversion $1\ mm = 0.039\ in$ (so the first datum changes to $(104)(0.039) = 4.06$). Use the converted data to construct a 90% confidence interval for the mean length of all adult male collared lemmings expressed in inches.
    c. Convert your answer in part (a) into inches directly and compare it to your answer in (b). This illustrates that if you construct a confidence interval in one system of units you can convert it directly into another system of units without having to convert all the data to the new units.

Answers

1. a. 98 ± 3.9
   b. 98 ± 5.2
2.
3. a. 386 ± 16.4
   b. 386 ± 33.6
4.
5. a. 933 ± 6.5
   b. 933 ± 8.5
   c. Asking for greater confidence requires a longer interval.
6.
7. 32.7 ± 2.9
8.
9. 47.3 ± 2.19
10.
11. 53.5 ± 0.56
12.
13. 8.2 ± 0.04
14.
15. 18.6 ± 1.4
16.
17. 9981 ± 486
18.
19. 104.998 ± 0.004
20.

21. a. 926 ± 200
    b. 419 ± 90
    c. 419 ± 91

7.3: Large Sample Estimation of a Population Proportion

Basic

1. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 90% confidence interval for the population proportion.
    a. $n = 25, \hat{p} = 0.7$
    b. $n = 50, \hat{p} = 0.7$
2. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 95% confidence interval for the population proportion.
    a. $n = 2500, \hat{p} = 0.22$
    b. $n = 1200, \hat{p} = 022$
3. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 98% confidence interval for the population proportion.
    a. $n = 80, \hat{p} = 0.4$
    b. $n = 325, \hat{p} = 0.4$
4. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 99.5% confidence interval for the population proportion.
    a. $n = 200, \hat{p} = 0.85$
    b. $n = 75, \hat{p} = 0.85$
5. In a random sample of size $1,100,338$ have the characteristic of interest.
    a. Compute the sample proportion $\hat{p}$ with the characteristic of interest.
    b. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion.
    c. Construct an 80% confidence interval for the population proportion p.
    d. Construct a 90% confidence interval for the population proportion p.
    e. Comment on why one interval is longer than the other.
6. In a random sample of size $2,400,420$ have the characteristic of interest.
    a. Compute the sample proportion $\hat{p}$ with the characteristic of interest.
    b. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion.
    c. Construct a 90% confidence interval for the population proportion p.
    d. Construct a 99% confidence interval for the population proportion p.
    e. Comment on why one interval is longer than the other.

Applications

Q7.3.7

A security feature on some web pages is graphic representations of words that are readable by human beings but not machines. When a certain design format was tested on 450 subjects, by having them attempt to read ten disguised words, 448 subjects could read all the words.

a. Give a point estimate of the proportion p of all people who could read words disguised in this way.
b. Show that the sample is not sufficiently large to construct a confidence interval for the proportion of all people who could read words disguised in this way.

Q7.3.8

In a random sample of 900 adults, 42 defined themselves as vegetarians.

a. Give a point estimate of the proportion of all adults who would define themselves as vegetarians.

b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
c. Construct an 80% confidence interval for the proportion of all adults who would define themselves as vegetarians.

Q7.3.9

In a random sample of 250 employed people, 61 said that they bring work home with them at least occasionally.

a. Give a point estimate of the proportion of all employed people who bring work home with them at least occasionally.
b. Construct a 99% confidence interval for that proportion.

Q7.3.10

In a random sample of $1,250$ household moves, 822 were moves to a location within the same county as the original residence.

a. Give a point estimate of the proportion of all household moves that are to a location within the same county as the original residence.
b. Construct a 98% confidence interval for that proportion.

Q7.3.11

In a random sample of $12,447$ hip replacement or revision surgery procedures nationwide, 162 patients developed a surgical site infection.

a. Give a point estimate of the proportion of all patients undergoing a hip surgery procedure who develop a surgical site infection.
b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
c. Construct a 95% confidence interval for the proportion of all patients undergoing a hip surgery procedure who develop a surgical site infection.

Q7.3.12

In a certain region prepackaged products labeled 500 g must contain on average at least 500 grams of the product, and at least 90% of all packages must weigh at least 490 grams. In a random sample of 300 packages, 288 weighed at least 490 grams.

a. Give a point estimate of the proportion of all packages that weigh at least 490 grams.
b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
c. Construct a 99.8% confidence interval for the proportion of all packages that weigh at least 490 grams.

Q7.3.13

A survey of 50 randomly selected adults in a small town asked them if their opinion on a proposed "no cruising" restriction late at night. Responses were coded 1 for in favor, 0 for indifferent, and 2 for opposed, with the results shown in the table.

$$\begin{array}{cccccccccc} 1 & 0 & 2 & 0 & 1 & 0 & 0 & 1 & 1 & 2 \\ 0 & 2 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 \\ 0 & 2 & 1 & 2 & 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 & 1 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 & 2 \end{array} \tag{7.E.7}$$

a. Give a point estimate of the proportion of all adults in the community who are indifferent concerning the proposed restriction.
b. Assuming that the sample is sufficiently large, construct a 90% confidence interval for the proportion of all adults in the community who are indifferent concerning the proposed restriction.

Q7.3.14

To try to understand the reason for returned goods, the manager of a store examines the records on 40 products that were returned in the last year. Reasons were coded by 1 for "defective," 2 for "unsatisfactory," and 0 for all other reasons, with the results shown in the table.

$$\begin{array}{cccccccccc} 0 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array} \tag{7.E.8}$$

a. Give a point estimate of the proportion of all returns that are because of something wrong with the product, that is, either defective or performed unsatisfactorily.
b. Assuming that the sample is sufficiently large, construct an 80% confidence interval for the proportion of all returns that are because of something wrong with the product.

Q7.3.15

In order to estimate the proportion of entering students who graduate within six years, the administration at a state university examined the records of 600 randomly selected students who entered the university six years ago, and found that 312 had graduated.

a. Give a point estimate of the six-year graduation rate, the proportion of entering students who graduate within six years.
b. Assuming that the sample is sufficiently large, construct a 98% confidence interval for the six-year graduation rate.

Q7.3.16

In a random sample of $2,300$ mortgages taken out in a certain region last year, 187 were adjustable-rate mortgages.

a. Give a point estimate of the proportion of all mortgages taken out in this region last year that were adjustable-rate mortgages.
b. Assuming that the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all mortgages taken out in this region last year that were adjustable-rate mortgages.

Q7.3.17

In a research study in cattle breeding, 159 of 273 cows in several herds that were in estrus were detected by means of an intensive once a day, one-hour observation of the herds in early morning.

a. Give a point estimate of the proportion of all cattle in estrus who are detected by this method.
b. Assuming that the sample is sufficiently large, construct a 90% confidence interval for the proportion of all cattle in estrus who are detected by this method.

Q7.3.18

A survey of $21,250$ households concerning telephone service gave the results shown in the table.

	Landline	No Landline
Cell phone	12,474	5,844
No cell phone	2,529	403

a. Give a point estimate for the proportion of all households in which there is a cell phone but no landline.
b. Assuming the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all households in which there is a cell phone but no landline.
c. Give a point estimate for the proportion of all households in which there is no telephone service of either kind.
d. Assuming the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all all households in which there is no telephone service of either kind.

Additional Exercises

19. In a random sample of 900 adults, 42 defined themselves as vegetarians. Of these 42, 29 were women.
    a. Give a point estimate of the proportion of all self-described vegetarians who are women.
    b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
    c. Construct a 90% confidence interval for the proportion of all all self-described vegetarians who are women.
20. A random sample of 185 college soccer players who had suffered injuries that resulted in loss of playing time was made with the results shown in the table. Injuries are classified according to severity of the injury and the condition under which it was sustained.

	Minor	Moderate	Serious
Practice	48	20	6
Game	62	32	17

a. Give a point estimate for the proportion p of all injuries to college soccer players that are sustained in practice.
b. Construct a 95% confidence interval for the proportion p of all injuries to college soccer players that are sustained in practice.
c. Give a point estimate for the proportion p of all injuries to college soccer players that are either moderate or serious.
d. Construct a 95% confidence interval for the proportion p of all injuries to college soccer players that are either moderate or serious.

21. The body mass index (BMI) was measured in $1,200$ randomly selected adults, with the results shown in the table.

	BMI		
	Under 18.5	18.5–25	Over 25
Men	36	165	315
Women	75	274	335

a. Give a point estimate for the proportion of all men whose BMI is over 25.
b. Assuming the sample is sufficiently large, construct a 99% confidence interval for the proportion of all men whose BMI is over 25.
c. Give a point estimate for the proportion of all adults, regardless of gender, whose BMI is over 25.
d. Assuming the sample is sufficiently large, construct a 99% confidence interval for the proportion of all adults, regardless of gender, whose BMI is over 25.

22. Confidence intervals constructed using the formula in this section often do not do as well as expected unless n is quite large, especially when the true population proportion is close to either 0 or 1. In such cases a better result is obtained by adding two successes and two failures to the actual data and then computing the confidence interval. This is the same as using the formula

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \quad \text{(7.E.9)}$$

where

$$\tilde{p} = \frac{x+2}{n+4} \text{ and } \tilde{n} = n+4$$

Suppose that in a random sample of 600 households, 12 had no telephone service of any kind. Use the adjusted confidence interval procedure just described to form a 99.9% confidence interval for the proportion of all households that have no telephone service of any kind.

Large Data Set Exercises

Large Data Set missing from the original

23. Large **Data Sets 4 and 4A** list the results of 500 tosses of a die. Let p denote the proportion of all tosses of this die that would result in a four. Use the sample data to construct a 90% confidence interval for p.
24. Large **Data Set 6** records results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate. Use the full data set (400 observations) to construct a 98% confidence interval for the proportion p of all voters who prefer Candidate A.
25. Lines 2 through 536 in **Data Set 11** is a sample of 535 real estate sales in a certain region in 2008. Those that were foreclosure sales are identified with a 1 in the second column.

a. Use these data to construct a point estimate $\hat{p}$ of the proportion p of all real estate sales in this region in 2008 that were foreclosure sales.

b. Use these data to construct a 90% confidence for p.

26. Lines 537 through 1106 in Large Data Set 11 is a sample of 570 real estate sales in a certain region in 2010. Those that were foreclosure sales are identified with a 1 in the second column.
    a. Use these data to construct a point estimate $\hat{p}$ of the proportion p of all real estate sales in this region in 2010 that were foreclosure sales.
    b. Use these data to construct a 90% confidence for p.

Answers

1. a. $(0.5492, 0.8508)$
   b. $(0.5934, 0.8066)$
2.
3. a. $(0.2726, 0.5274)$
   b. $(0.3368, 0.4632)$
4.
5. a. 0.3073
   b. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.31 \pm 0.04$ and $[0.27, 0.35] \subset [0, 1]$
   c. $(0.2895, 0.3251)$
   d. $(0.2844, 0.3302)$
   e. Asking for greater confidence requires a longer interval.
6.
7. a. 0.9956
   b. $(0.9862, 1.005)$
8.
9. a. 0.244
   b. $(0.1740, 0.3140)$
10.
11. a. 0.013
    b. $(0.01, 0.016)$
    c. $(0.011, 0.015)$
12.
13. a. 0.52
    b. $(0.4038, 0.6362)$
14.
15. a. 0.52
    b. $(0.4726, 0.5674)$
16.
17. a. 0.5824
    b. $(0.5333, 0.6315)$
18.
19. a. 0.69
    b. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.69 \pm 0.21$ and $[0.48, 0.90] \subset [0, 1]$
    c. 0.69 ± 0.12
20.
21. a. 0.6105
    b. $(0.5552, 0.6658)$
    c. 0.5583

d. $(0.5214, 0.5952)$

22.

23. $(0.1368, 0.1912)$

24.

25. a. $\hat{p} = 0.2280$
    b. $(0.1982, 0.2579)$

7.4: Sample Size Considerations

Basic

1. Estimate the minimum sample size needed to form a confidence interval for the mean of a population having the standard deviation shown, meeting the criteria given.
   a. $\sigma = 30, 95\%$ confidence, $E = 10$
   b. $\sigma = 30, 99\%$ confidence, $E = 10$
   c. $\sigma = 30, 95\%$ confidence, $E = 5$
2. Estimate the minimum sample size needed to form a confidence interval for the mean of a population having the standard deviation shown, meeting the criteria given.
   a. $\sigma = 4, 95\%$ confidence, $E = 1$
   b. $\sigma = 4, 99\%$ confidence, $E = 1$
   c. $\sigma = 4, 95\%$ confidence, $E = 0.5$
3. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. $p \approx 0.37, 80\%$ confidence, $E = 0.05$
   b. $p \approx 0.37, 90\%$ confidence, $E = 0.05$
   c. $p \approx 0.37, 80\%$ confidence, $E = 0.01$
4. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. $p \approx 0.81, 95\%$ confidence, $E = 0.02$
   b. $p \approx 0.81, 99\%$ confidence, $E = 0.02$
   c. $p \approx 0.81, 95\%$ confidence, $E = 0.01$
5. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. 80% confidence, $E = 0.05$
   b. 90% confidence, $E = 0.05$
   c. 80% confidence, $E = 0.01$
6. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. 95% confidence, $E = 0.02$
   b. 99% confidence, $E = 0.02$
   c. 95% confidence, $E = 0.01$

Applications

7. A software engineer wishes to estimate, to within 5 seconds, the mean time that a new application takes to start up, with 95% confidence. Estimate the minimum size sample required if the standard deviation of start up times for similar software is 12 seconds.
8. A real estate agent wishes to estimate, to within $\$2.50$, the mean retail cost per square foot of newly built homes, with 80% confidence. He estimates the standard deviation of such costs at $\$5.00$. Estimate the minimum size sample required.
9. An economist wishes to estimate, to within 2 minutes, the mean time that employed persons spend commuting each day, with 95% confidence. On the assumption that the standard deviation of commuting times is 8 minutes, estimate the minimum size

sample required.

10. A motor club wishes to estimate, to within 1 cent, the mean price of 1 gallon of regular gasoline in a certain region, with 98% confidence. Historically the variability of prices is measured by $\sigma = \$0.03$Estimate the minimum size sample required.
11. A bank wishes to estimate, to within $\$25$, the mean average monthly balance in its checking accounts, with 99.8% confidence. Assuming $\sigma = \$250$, estimate the minimum size sample required.
12. A retailer wishes to estimate, to within 15 seconds, the mean duration of telephone orders taken at its call center, with 99.5% confidence. In the past the standard deviation of call length has been about 1.25 minutes. Estimate the minimum size sample required. (Be careful to express all the information in the same units.)
13. The administration at a college wishes to estimate, to within two percentage points, the proportion of all its entering freshmen who graduate within four years, with 90% confidence. Estimate the minimum size sample required.
14. A chain of automotive repair stores wishes to estimate, to within five percentage points, the proportion of all passenger vehicles in operation that are at least five years old, with 98% confidence. Estimate the minimum size sample required.
15. An internet service provider wishes to estimate, to within one percentage point, the current proportion of all email that is spam, with 99.9% confidence. Last year the proportion that was spam was 71%. Estimate the minimum size sample required.
16. An agronomist wishes to estimate, to within one percentage point, the proportion of a new variety of seed that will germinate when planted, with 95% confidence. A typical germination rate is 97%. Estimate the minimum size sample required.
17. A charitable organization wishes to estimate, to within half a percentage point, the proportion of all telephone solicitations to its donors that result in a gift, with 90% confidence. Estimate the minimum sample size required, using the information that in the past the response rate has been about 30%.
18. A government agency wishes to estimate the proportion of drivers aged $16 - 24$ who have been involved in a traffic accident in the last year. It wishes to make the estimate to within one percentage point and at 90% confidence. Find the minimum sample size required, using the information that several years ago the proportion was 0.12.

Additional Exercises

19. An economist wishes to estimate, to within six months, the mean time between sales of existing homes, with 95% confidence. Estimate the minimum size sample required. In his experience virtually all houses are re-sold within 40 months, so using the Empirical Rule he will estimate σ by one-sixth the range, or $40/6 = 6.7$.
20. A wildlife manager wishes to estimate the mean length of fish in a large lake, to within one inch, with 80% confidence. Estimate the minimum size sample required. In his experience virtually no fish caught in the lake is over 23 inches long, so using the Empirical Rule he will estimate σ by one-sixth the range, or $23/6 = 3.8$.
21. You wish to estimate the current mean birth weight of all newborns in a certain region, to within 1 ounce ($1/16$ pound) and with 95% confidence. A sample will cost $\$400$ plus $\$1.50$ for every newborn weighed. You believe the standard deviations of weight to be no more than 1.25 pounds. You have $\$2,500$ to spend on the study.
    a. Can you afford the sample required?
    b. If not, what are your options?
22. You wish to estimate a population proportion to within three percentage points, at 95% confidence. A sample will cost $\$500$ plus 50 cents for every sample element measured. You have $\$1,000$ to spend on the study.
    a. Can you afford the sample required?
    b. If not, what are your options?

Answers

1. a. 35
   b. 60
   c. 139
2.
3. a. 154
   b. 253
   c. 3832
4.

5. a. **165**
   b. **271**
   c. **4109**
6.
7. **23**
8.
9. **62**
10.
11. **955**
12.
13. **1692**
14.
15. **22, 301**
16.
17. **22, 731**
18.
19. **5**
20.
21. a. no
    b. decrease the confidence level

This page titled 7.E: Estimation (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **7.E: Estimation (Exercises)** has no license indicated.

CHAPTER OVERVIEW

8: Testing Hypotheses

In the sampling that we have studied so far the goal has been to estimate a population parameter. But the sampling done by the government agency has a somewhat different objective, not so much to *estimate* the population mean μ as to *test* an assertion—or a hypothesis—about it, namely, whether it is as large as 75 or not. The agency is not necessarily interested in the actual value of μ, just whether it is as claimed. Their sampling is done to perform a test of hypotheses, the subject of this chapter.

8.1: The Elements of Hypothesis Testing

8.2: Tests for a Population Mean

8.2.1: The Observed Significance of a Test

8.2.2: Small Sample Tests for a Population Mean

8.3: Tests for a Population Proportion

8.E: Testing Hypotheses (Exercises)

This page titled 8: Testing Hypotheses is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

8.1: The Elements of Hypothesis Testing

Learning Objectives

- To understand the logical framework of tests of hypotheses.
- To learn basic terminology connected with hypothesis testing.
- To learn fundamental facts about hypothesis testing.

Types of Hypotheses

A *hypothesis* about the value of a population parameter is an assertion about its value. As in the introductory example we will be concerned with testing the truth of two competing hypotheses, only one of which can be true.

Definition: null hypothesis and alternative hypothesis

- The *null hypothesis*, denoted H_0, is the statement about the population parameter that is assumed to be true unless there is convincing evidence to the contrary.
- The *alternative hypothesis*, denoted H_a, is a statement about the population parameter that is contradictory to the null hypothesis, and is accepted as true only if there is convincing evidence in favor of it.

Definition: statistical procedure

Hypothesis testing is a *statistical procedure* in which a choice is made between a null hypothesis and an alternative hypothesis based on information in a sample.

The end result of a hypotheses testing procedure is a choice of one of the following two possible conclusions:

1. Reject H_0 (and therefore accept H_a), or
2. Fail to reject H_0 (and therefore fail to accept H_a).

The null hypothesis typically represents the status quo, or what has historically been true. In the example of the respirators, we would believe the claim of the manufacturer unless there is reason not to do so, so the null hypotheses is $H_0 : \mu = 75$. The alternative hypothesis in the example is the contradictory statement $H_a : \mu < 75$. The null hypothesis will always be an assertion containing an equals sign, but depending on the situation the alternative hypothesis can have any one of three forms: with the symbol $<$, as in the example just discussed, with the symbol $>$, or with the symbol $\neq$. The following two examples illustrate the latter two cases.

Example 8.1.1

A publisher of college textbooks claims that the average price of all hardbound college textbooks is $\$127.50$. A student group believes that the actual mean is higher and wishes to test their belief. State the relevant null and alternative hypotheses.

Solution

The default option is to accept the publisher's claim unless there is compelling evidence to the contrary. Thus the null hypothesis is $H_0 : \mu = 127.50$. Since the student group thinks that the average textbook price is greater than the publisher's figure, the alternative hypothesis in this situation is $H_a : \mu > 127.50$.

Example 8.1.2

The recipe for a bakery item is designed to result in a product that contains 8 grams of fat per serving. The quality control department samples the product periodically to insure that the production process is working as designed. State the relevant null and alternative hypotheses.

Solution

The default option is to assume that the product contains the amount of fat it was formulated to contain unless there is compelling evidence to the contrary. Thus the null hypothesis is $H_0 : \mu = 8.0$. Since to contain either more fat than desired or to contain less fat than desired are both an indication of a faulty production process, the alternative hypothesis in this situation is that the mean is different from 8.0, so $H_a : \mu \neq 8.0$.

In Example 8.1.1, the textbook example, it might seem more natural that the publisher's claim be that the average price is at most $\$127.50$, not exactly $\$127.50$. If the claim were made this way, then the null hypothesis would be $H_0 : \mu \leq 127.50$, and the value $\$127.50$ given in the example would be the one that is least favorable to the publisher's claim, the null hypothesis. It is always true that if the null hypothesis is retained for its least favorable value, then it is retained for every other value.

Thus in order to make the null and alternative hypotheses easy for the student to distinguish, in every example and problem in this text we will always present one of the two competing claims about the value of a parameter with an equality. The claim expressed with an equality is the null hypothesis. This is the same as always stating the null hypothesis in the least favorable light. So in the introductory example about the respirators, we stated the manufacturer's claim as "the average is 75 minutes" instead of the perhaps more natural "the average is at least 75 minutes," essentially reducing the presentation of the null hypothesis to its worst case.

The first step in hypothesis testing is to identify the null and alternative hypotheses.

The Logic of Hypothesis Testing

Although we will study hypothesis testing in situations other than for a single population mean (for example, for a population proportion instead of a mean or in comparing the means of two different populations), in this section the discussion will always be given in terms of a single population mean μ.

The null hypothesis always has the form $H_0 : \mu = \mu_0$ for a specific number μ_0 (in the respirator example $\mu_0 = 75$, in the textbook example $\mu_0 = 127.50$, and in the baked goods example $\mu_0 = 8.0$). Since the null hypothesis is accepted unless there is strong evidence to the contrary, the test procedure is based on the initial assumption that H_0 is true. This point is so important that we will repeat it in a display:

> The test procedure is based on the initial assumption that H_0 is true.

The criterion for judging between H_0 and H_a based on the sample data is: if the value of $\overline{X}$ would be highly unlikely to occur if H_0 were true, but favors the truth of H_a, then we reject H_0 in favor of H_a. Otherwise we do not reject H_0.

Supposing for now that $\overline{X}$ follows a normal distribution, when the null hypothesis is true the density function for the sample mean $\overline{X}$ must be as in Figure 8.1.1: a bell curve centered at μ_0. Thus if H_0 is true then $\overline{X}$ is likely to take a value near μ_0 and is unlikely to take values far away. Our decision procedure therefore reduces simply to:

- if H_a has the form $H_a : \mu < \mu_0$ then reject H_0 if $\bar{x}$ is far to the left of μ_0;
- if H_a has the form $H_a : \mu > \mu_0$ then reject H_0 if $\bar{x}$ is far to the right of μ_0;
- if H_a has the form $H_a : \mu \neq \mu_0$ then reject H_0 if $\bar{x}$ is far away from μ_0 in either direction.

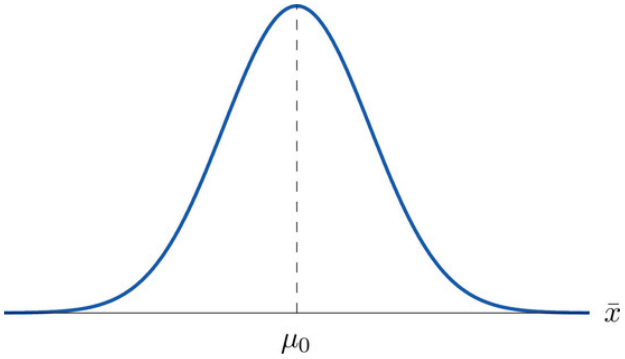


Figure 8.1.1 : The Density Curve for $\overline{X}$ *if* H_0 *Is True*

Think of the respirator example, for which the null hypothesis is $H_0 : \mu = 75$, the claim that the average time air is delivered for all respirators is 75 minutes. If the sample mean is 75 or greater then we certainly would not reject H_0 (since there is no issue with an emergency respirator delivering air even longer than claimed).

If the sample mean is slightly less than 75 then we would logically attribute the difference to sampling error and also not reject H_0 either.

Values of the sample mean that are smaller and smaller are less and less likely to come from a population for which the population mean is 75. Thus if the sample mean is far less than 75, say around 60 minutes or less, then we would certainly reject H_0, because we know that it is highly unlikely that the average of a sample would be so low if the population mean were 75. This is the rare event criterion for rejection: what we actually observed ($\overline{X} < 60$) would be so rare an event if $\mu = 75$ were true that we regard it as much more likely that the alternative hypothesis $\mu < 75$ holds.

In summary, to decide between H_0 and H_a in this example we would select a "rejection region" of values sufficiently far to the left of 75, based on the rare event criterion, and reject H_0 if the sample mean $\overline{X}$ lies in the rejection region, but not reject H_0 if it does not.

The Rejection Region

Each different form of the alternative hypothesis Ha has its own kind of rejection region:

- if (as in the respirator example) H_a has the form $H_a : \mu < \mu_0$, we reject H_0 if $\bar{x}$ is far to the left of μ_0, that is, to the left of some number C, so the rejection region has the form of an interval $(-\infty, C]$;
- if (as in the textbook example) H_a has the form $H_a : \mu > \mu_0$, we reject H_0 if $\bar{x}$ is far to the right of μ_0, that is, to the right of some number C, so the rejection region has the form of an interval $[C, \infty)$;
- if (as in the baked good example) H_a has the form $H_a : \mu \neq \mu_0$, we reject H_0 if $\bar{x}$ is far away from μ_0 in either direction, that is, either to the left of some number C or to the right of some other number C', so the rejection region has the form of the union of two intervals $(-\infty, C] \cup [C', \infty)$.

The key issue in our line of reasoning is the question of how to determine the number C or numbers C and C', called the critical value or critical values of the statistic, that determine the rejection region.

Definition: critical values

The critical value or critical values of a test of hypotheses are the number or numbers that determine the rejection region.

Suppose the rejection region is a single interval, so we need to select a single number C. Here is the procedure for doing so. We select a small probability, denoted α, say 1%, which we take as our definition of "rare event:" an event is "rare" if its probability of occurrence is less than α. (In all the examples and problems in this text the value of α will be given already.) The probability that $\overline{X}$ takes a value in an interval is the area under its density curve and above that interval, so as shown in Figure 8.1.2 (drawn under the assumption that H_0 is true, so that the curve centers at μ_0) the critical value C is the value of $\overline{X}$ that cuts off a tail area α in the probability density curve of $\overline{X}$. When the rejection region is in two pieces, that is, composed of two intervals, the total area above both of them must be α, so the area above each one is $\alpha/2$, as also shown in Figure 8.1.2.

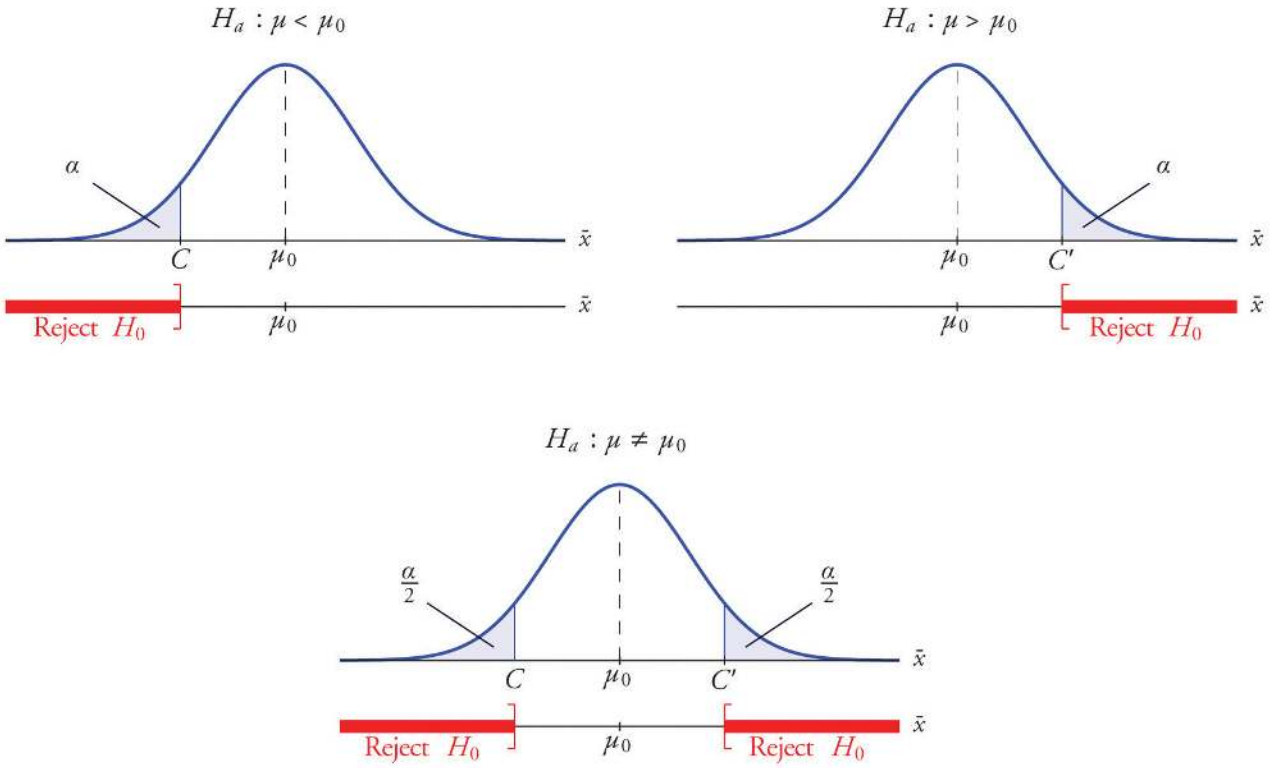


Figure 8.1.2

The number α is the total area of a tail or a pair of tails.

Example 8.1.3

In the context of Example 8.1.2, suppose that it is known that the population is normally distributed with standard deviation $\alpha = 0.15$ gram, and suppose that the test of hypotheses $H_0 : \mu = 8.0$ versus $H_a : \mu \neq 8.0$ will be performed with a sample of size 5. Construct the rejection region for the test for the choice $\alpha = 0.10$. Explain the decision procedure and interpret it.

Solution

If H_0 is true then the sample mean $\overline{X}$ is normally distributed with mean and standard deviation

$$\begin{aligned} \mu_{\overline{X}} &= \mu \\ &= 8.0 \end{aligned}$$

$$\begin{aligned} \sigma_{\overline{X}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{0.15}{\sqrt{5}} \\ &= 0.067 \end{aligned}$$

Since H_a contains the $\neq$ symbol the rejection region will be in two pieces, each one corresponding to a tail of area $\alpha/2 = 0.10/2 = 0.05$. From Figure 7.1.6, $z_{0.05} = 1.645$, so C and C' are 1.645 standard deviations of $\overline{X}$ to the right and left of its mean 8.0:

$$C = 8.0 - (1.645)(0.067) = 7.89 \text{ and } C' = 8.0 + (1.645)(0.067) = 8.11$$

The result is shown in Figure 8.1.3.

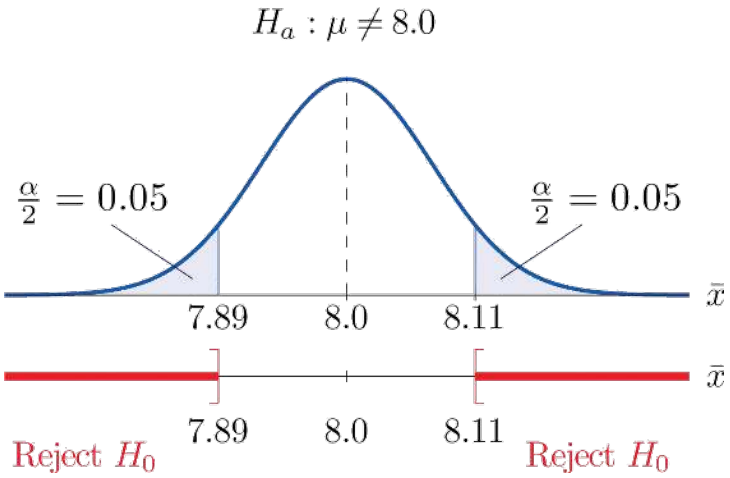


Figure **8.1.3**: Rejection Region for the Choice

The decision procedure is: take a sample of size **5** and compute the sample mean $\bar{x}$. If $\bar{x}$ is either **7.89** grams or less or **8.11** grams or more then reject the hypothesis that the average amount of fat in all servings of the product is **8.0** grams in favor of the alternative that it is different from **8.0** grams. Otherwise do not reject the hypothesis that the average amount is **8.0** grams.

The reasoning is that if the true average amount of fat per serving were **8.0** grams then there would be less than a **10%** chance that a sample of size **5** would produce a mean of either **7.89** grams or less or **8.11** grams or more. Hence if that happened it would be more likely that the value **8.0** is incorrect (always assuming that the population standard deviation is **0.15** gram).

Because the rejection regions are computed based on areas in tails of distributions, as shown in Figure **8.1.2**, hypothesis tests are classified according to the form of the alternative hypothesis in the following way.

Definitions: Test classifications

- If H_a has the form $\mu \neq \mu_0$ the test is called a **two-tailed test**.
- If H_a has the form $\mu < \mu_0$ the test is called a **left-tailed test**.
- If H_a has the form $\mu > \mu_0$the test is called a **right-tailed test**.

Each of the last two forms is also called a **one-tailed test**.

Two Types of Errors

The format of the testing procedure in general terms is to take a sample and use the information it contains to come to a decision about the two hypotheses. As stated before our decision will always be either

1. reject the null hypothesis H_0 in favor of the alternative H_a presented, or
2. do not reject the null hypothesis H_0 in favor of the alternative H_0 presented.

There are four possible outcomes of hypothesis testing procedure, as shown in the following table:

		True State of Nature	
	H_0 is true	H_0 is false	
Our Decision	Do not reject H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

As the table shows, there are two ways to be right and two ways to be wrong. Typically to reject H_0 when it is actually true is a more serious error than to fail to reject it when it is false, so the former error is labeled "**Type I**" and the latter error "**Type II**".

Definition: Type I and Type II errors

In a test of hypotheses:

- A *Type I error* is the decision to reject H_0 when it is in fact true.
- A *Type II error* is the decision not to reject H_0 when it is in fact not true.

Unless we perform a census we do not have certain knowledge, so we do not know whether our decision matches the true state of nature or if we have made an error. We reject H_0 if what we observe would be a "rare" event if H_0 were true. But rare events are not impossible: they occur with probability α. Thus when H_0 is true, a rare event will be observed in the proportion α of repeated similar tests, and H_0 will be erroneously rejected in those tests. Thus α is the probability that in following the testing procedure to decide between H_0 and H_a we will make a Type I error.

Definition: level of significance

The number α that is used to determine the rejection region is called the level of significance of the test. It is the probability that the test procedure will result in a *Type I error*.

The probability of making a Type II error is too complicated to discuss in a beginning text, so we will say no more about it than this: for a fixed sample size, choosing $alpha$ smaller in order to reduce the chance of making a Type I error has the effect of increasing the chance of making a **Type II error**. The only way to simultaneously reduce the chances of making either kind of error is to increase the sample size.

Standardizing the Test Statistic

Hypotheses testing will be considered in a number of contexts, and great unification as well as simplification results when the relevant sample statistic is *standardized* by subtracting its mean from it and then dividing by its standard deviation. The resulting statistic is called a *standardized test statistic*. In every situation treated in this and the following two chapters the standardized test statistic will have either the standard normal distribution or Student's t-distribution.

Definition: hypothesis test

A standardized test statistic for a hypothesis test is the statistic that is formed by subtracting from the statistic of interest its mean and dividing by its standard deviation.

For example, reviewing Example 8.1.3, if instead of working with the sample mean $\overline{X}$ we instead work with the test statistic

$$\frac{\overline{X} - 8.0}{0.067}$$

then the distribution involved is standard normal and the critical values are just $\pm z_{0.05}$. The extra work that was done to find that $C = 7.89$ and $C' = 8.11$ is eliminated. In every hypothesis test in this book the standardized test statistic will be governed by either the standard normal distribution or Student's t-distribution. Information about rejection regions is summarized in the following tables:

Table 8.1.1: When the test statistic has the standard normal distribution

Symbol in H_a	Terminology	Rejection Region
$<$	Left-tailed test	$(-\infty, -z_\alpha]$
$>$	Right-tailed test	$[z_\alpha, \infty)$
$\neq$	Two-tailed test	$(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$

Table 8.1.2: When the test statistic has Student's t-distribution

Symbol in H_a	Terminology	Rejection Region
$<$	Left-tailed test	$(-\infty, -t_\alpha]$
$>$	Right-tailed test	$[t_\alpha, \infty)$
$\neq$	Two-tailed test	$(-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty)$

Every instance of hypothesis testing discussed in this and the following two chapters will have a rejection region like one of the six forms tabulated in the tables above.

No matter what the context a test of hypotheses can always be performed by applying the following systematic procedure, which will be illustrated in the examples in the succeeding sections.

Systematic Hypothesis Testing Procedure: Critical Value Approach

1. Identify the null and alternative hypotheses.
2. Identify the relevant test statistic and its distribution.
3. Compute from the data the value of the test statistic.
4. Construct the rejection region.
5. Compare the value computed in Step 3 to the rejection region constructed in Step 4 and make a decision. Formulate the decision in the context of the problem, if applicable.

The procedure that we have outlined in this section is called the “Critical Value Approach” to hypothesis testing to distinguish it from an alternative but equivalent approach that will be introduced at the end of Section 8.3.

Key Takeaway

- A test of hypotheses is a statistical process for deciding between two competing assertions about a population parameter.
- The testing procedure is formalized in a five-step procedure.

This page titled 8.1: The Elements of Hypothesis Testing is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **8.1: The Elements of Hypothesis Testing** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

8.2: Tests for a Population Mean

Learning Objectives

- To learn how to apply the five-step test procedure for a test of hypotheses concerning a population mean when the sample size is large.
- To learn how to interpret the result of a test of hypotheses in the context of the original narrated situation.

In this section we describe and demonstrate the procedure for conducting a test of hypotheses about the mean of a population in the case that the sample size n is at least 30. The Central Limit Theorem states that $\overline{X}$ is approximately normally distributed, and has mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, where μ and σ are the mean and the standard deviation of the population. This implies that the statistic

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution, which means that probabilities related to it are given in Figure 7.1.5 and the last line in Figure 7.1.6.

If we know σ then the statistic in the display is our test statistic. If, as is typically the case, we do not know σ, then we replace it by the sample standard deviation s. Since the sample is large the resulting test statistic still has a distribution that is approximately standard normal.

Standardized Test Statistics for Large Sample Hypothesis Tests Concerning a Single Population Mean

- If σ is known: $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
- If σ is unknown: $Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

The test statistic has the standard normal distribution.

The distribution of the standardized test statistic and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure 8.2.1.

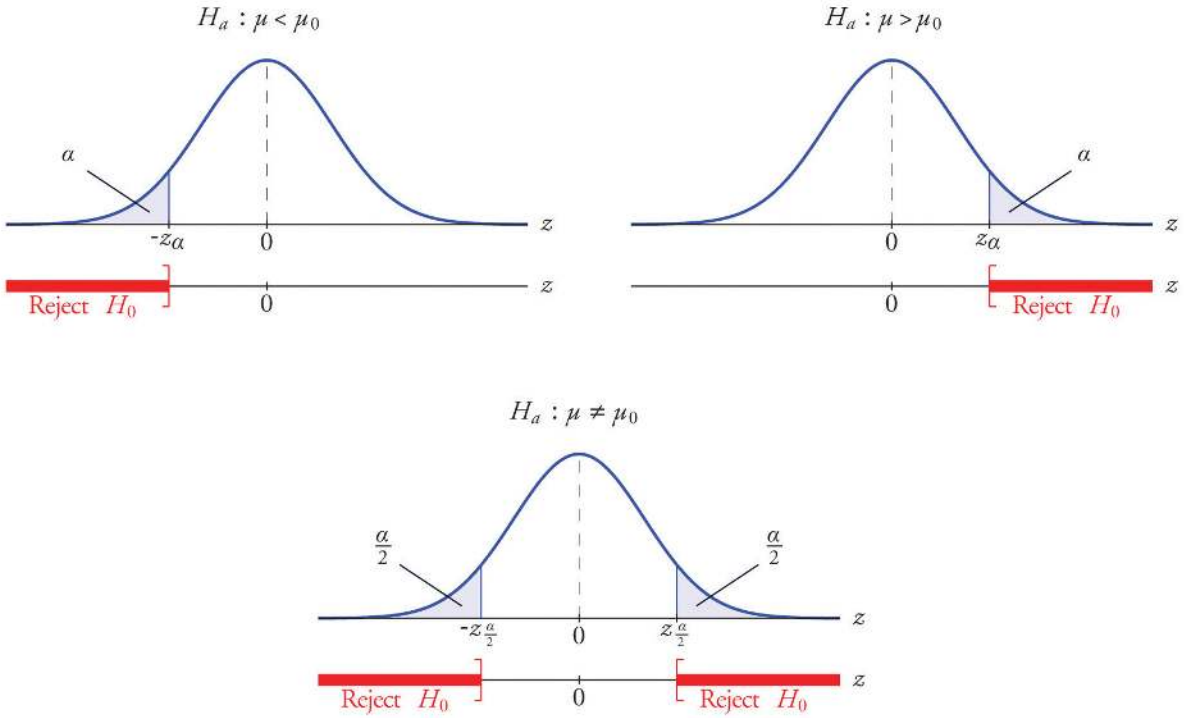


Figure 8.2.1: Distribution of the Standardized Test Statistic and the Rejection Region

Example 8.2.1

It is hoped that a newly developed pain reliever will more quickly produce perceptible reduction in pain to patients after minor surgeries than a standard pain reliever. The standard pain reliever is known to bring relief in an average of 3.5 minutes with standard deviation 2.1 minutes. To test whether the new pain reliever works more quickly than the standard one, 50 patients with minor surgeries were given the new pain reliever and their times to relief were recorded. The experiment yielded sample mean $\bar{x} = 3.1$ minutes and sample standard deviation $s = 1.5$ minutes. Is there sufficient evidence in the sample to indicate, at the 5% level of significance, that the newly developed pain reliever does deliver perceptible relief more quickly?

Solution

We perform the test of hypotheses using the five-step procedure given at the end of Section 8.1.

- **Step 1**. The natural assumption is that the new drug is no better than the old one, but must be proved to be better. Thus if μ denotes the average time until all patients who are given the new drug experience pain relief, the hypothesis test is

$$H_0 : \mu = 3.5$$
$$\text{vs}$$
$$H_a : \mu < 3.5 \text{ @ } \alpha = 0.05$$

- **Step 2**. The sample is large, but the population standard deviation is unknown (the 2.1 minutes pertains to the old drug, not the new one). Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  and has the standard normal distribution.
- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.1 - 3.5}{1.5/\sqrt{50}} = -1.886$$

- **Step 4**. Since the symbol in H_a is "<" this is a left-tailed test, so there is a single critical value, $-z_\alpha = -z_{0.005}$, which from the last line in Figure 7.1.6 we read off as -1.645. The rejection region is $(-\infty, -1.645]$.

- **Step 5**. As shown in Figure 8.2.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the average time until patients experience perceptible relief from pain using the new pain reliever is smaller than the average time for the standard pain reliever.

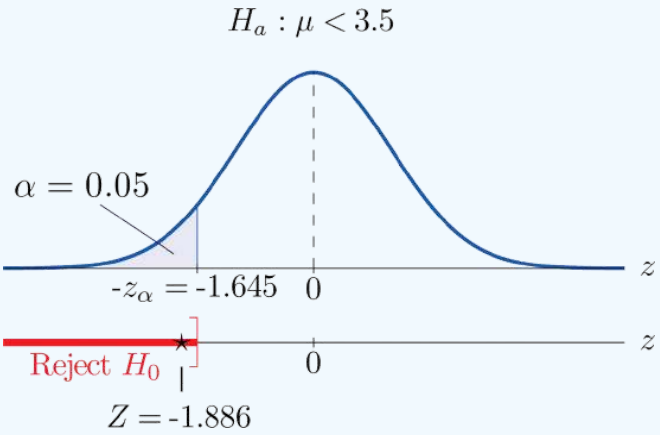


Figure 8.2.2: Rejection Region and Test Statistic for "Example 8.2.1"

Example 8.2.2

A cosmetics company fills its best-selling 8 ounce jars of facial cream by an automatic dispensing machine. The machine is set to dispense a mean of 8.1 ounces per jar. Uncontrollable factors in the process can shift the mean away from 8.1 and cause either underfill or overfill, both of which are undesirable. In such a case the dispensing machine is stopped and recalibrated. Regardless of the mean amount dispensed, the standard deviation of the amount dispensed always has value 0.22 ounce. A quality control engineer routinely selects 30 jars from the assembly line to check the amounts filled. On one occasion, the sample mean is $\bar{x} = 8.2$ ounces and the sample standard deviation is $s = 0.25$ ounce. Determine if there is sufficient evidence in the sample to indicate, at the 1% level of significance, that the machine should be recalibrated.

Solution

- **Step 1**. The natural assumption is that the machine is working properly. Thus if μ denotes the mean amount of facial cream being dispensed, the hypothesis test is

$$H_0 : \mu = 8.1$$
$$\text{vs}$$
$$H_a : \mu \neq 8.1 \text{ @ } \alpha = 0.01$$

- **Step 2**. The sample is large and the population standard deviation is known. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

  and has the standard normal distribution.
- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{8.2 - 8.1}{0.22/\sqrt{30}} = 2.490$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm z_{\alpha/2} = \pm z_{0.005}$, which from the last line in Figure 7.1.6 "Critical Values of " we read off as ± 2.576. The rejection region is $(-\infty, -2.576] \cup [2.576, \infty)$.
- **Step 5**. As shown in Figure 8.2.3 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the average amount of product dispensed is different from 8.1 ounce. We conclude that the machine does not need to be recalibrated.

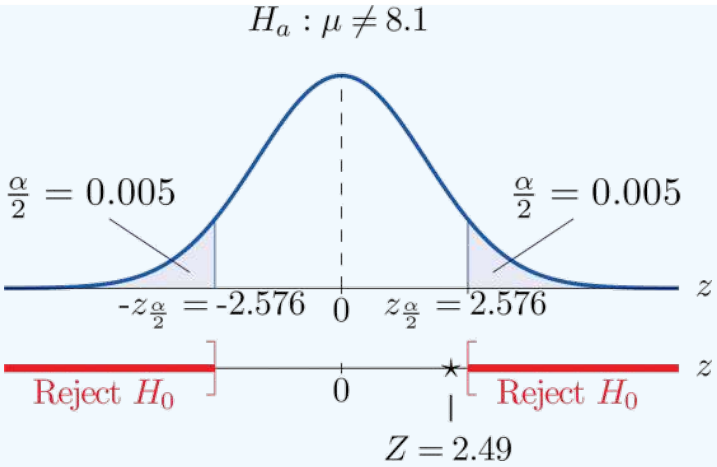


Figure **8.2.3**: Rejection Region and Test Statistic for "Example **8.2.2**"

Key Takeaway

- There are two formulas for the test statistic in testing hypotheses about a population mean with large samples. Both test statistics follow the standard normal distribution.
- The population standard deviation is used if it is known, otherwise the sample standard deviation is used.
- The same five-step procedure is used with either test statistic.

This page titled 8.2: Tests for a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **8.2: Large Sample Tests for a Population Mean** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

8.2.1: The Observed Significance of a Test

Learning Objectives

- To learn what the observed significance of a test is.
- To learn how to compute the observed significance of a test.
- To learn how to apply the p-value approach to hypothesis testing.

The Observed Significance

The conceptual basis of our testing procedure is that we reject H_0 only if the data that we obtained would constitute a rare event if H_0 were actually true. The level of significance α specifies what is meant by "rare." The observed significance of the test is a measure of how rare the value of the test statistic that we have just observed would be if the null hypothesis were true. That is, the observed significance of the test just performed is the probability that, if the test were repeated with a new sample, the result of the new test would be at least as contrary to H_0 and in support of H_a as what was observed in the original test.

Definition: observed significance

The *observed significance* or p-value of a specific test of hypotheses is the probability, on the supposition that H_0 is true, of obtaining a result at least as contrary to H_0 and in favor of H_a as the result actually observed in the sample data.

Think back to "Example 8.2.1", Section 8.2 concerning the effectiveness of a new pain reliever. This was a left-tailed test in which the value of the test statistic was -1.886. To be as contrary to H_0 and in support of H_a as the result $Z = -1.886$ actually observed means to obtain a value of the test statistic in the interval $(-\infty, -1.886]$. Rounding -1.886 to -1.89, we can read directly from Figure 7.1.5 that $P(Z \leq -1.89) = 0.0294$. Thus the p-value or observed significance of the test in "Example 8.2.1", Section 8.2 is 0.0294 or about 3%. Under repeated sampling from this population, if H_0 were true then only about 3% of all samples of size 50 would give a result as contrary to H_0 and in favor of H_a as the sample we observed. Note that the probability 0.0294 is the area of the left tail cut off by the test statistic in this left-tailed test.

Analogous reasoning applies to a right-tailed or a two-tailed test, except that in the case of a two-tailed test being as far from 0 as the observed value of the test statistic but on the opposite side of 0 is just as contrary to H_0 as being the same distance away and on the same side of 0, hence the corresponding tail area is doubled.

Computational Definition of the Observed Significance of a Test of Hypotheses

The **observed significance** of a test of hypotheses is the area of the tail of the distribution cut off by the test statistic (times two in the case of a two-tailed test).

Example 8.2.1.1

Compute the observed significance of the test performed in "Example 8.2.2", Section 8.2.

Solution

The value of the test statistic was $z = 2.490$, which by Figure 7.1.5 cuts off a tail of area 0.0064, as shown in Figure $8.2.1.1$. Since the test was two-tailed, the observed significance is $2 \times 0.0064 = 0.0128$.

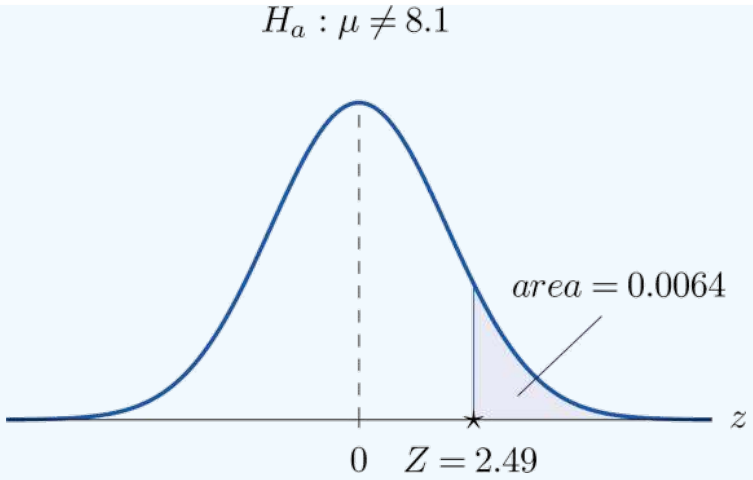


Figure **8.2.1.1**: Area of the Tail for Example **8.2.1.1**.

The p-value Approach to Hypothesis Testing

In "Example 8.2.1", Section 8.2 the test was performed at the 5% level of significance: the definition of "rare" event was probability $\alpha = 0.05$ or less. We saw above that the observed significance of the test was $p = 0.0294$ or about 3%. Since $p = 0.0294 < 0.05 = \alpha$ (or 3% is less than 5%), the decision turned out to be to reject: what was observed was sufficiently unlikely to qualify as an event so rare as to be regarded as (practically) incompatible with H_0.

In "Example 8.2.2", Section 8.2 the test was performed at the 1% level of significance: the definition of "rare" event was probability $\alpha = 0.01$ or less. The observed significance of the test was computed in "Example **8.2.1.1**" as $p = 0.0128$ or about 1.3%. Since $p = 0.0128 > 0.01 = \alpha$ (or 1.3% is greater than 1%), the decision turned out to be not to reject. The event observed was unlikely, but not sufficiently unlikely to lead to rejection of the null hypothesis.

The reasoning just presented is the basis for a slightly different but equivalent formulation of the hypothesis testing process. The first three steps are the same as before, but instead of using α to compute critical values and construct a rejection region, one computes the p-value p of the test and compares it to α, rejecting H_0 if $p \leq \alpha$ and not rejecting if $p > \alpha$.

Systematic Hypothesis Testing Procedure: *p*-Value Approach

1. Identify the null and alternative hypotheses.
2. Identify the relevant test statistic and its distribution.
3. Compute from the data the value of the test statistic.
4. Compute the p-value of the test.
5. Compare the value computed in Step 4 to significance level α and make a decision: reject H_0 if $p \leq \alpha$ and do not reject H_0 if $p > \alpha$. Formulate the decision in the context of the problem, if applicable.

Example 8.2.1.2

The total score in a professional basketball game is the sum of the scores of the two teams. An expert commentator claims that the average total score for NBA games is 202.5. A fan suspects that this is an overstatement and that the actual average is less than 202.5. He selects a random sample of 85 games and obtains a mean total score of 199.2 with standard deviation 19.63. Determine, at the 5% level of significance, whether there is sufficient evidence in the sample to reject the expert commentator's claim.

Solution

- **Step 1**. Let μ be the true average total game score of all NBA games. The relevant test is

$$H_0 : \mu = 202.5$$
$$\text{vs}$$
$$H_a : \mu < 202.5 \ @ \ \alpha = 0.05$$

- **Step 2**. The sample is large and the population standard deviation is unknown. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the standard normal distribution.

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{199.2 - 202.5}{19.63/\sqrt{85}} = -1.55$$

- **Step 4**. The area of the left tail cut off by $z = -1.55$ is, by Figure 7.1.5, 0.0606, as illustrated in Figure 8.2.1.2. Since the test is left-tailed, the p-value is just this number, $p = 0.0606$.
- **Step 5**. Since $p = 0.0606 > 0.05 = \alpha$, the decision is not to reject H_0. In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 5% level of significance, to conclude that the average total score of NBA games is less than 202.5.

$$H_a : \mu < 202.5$$

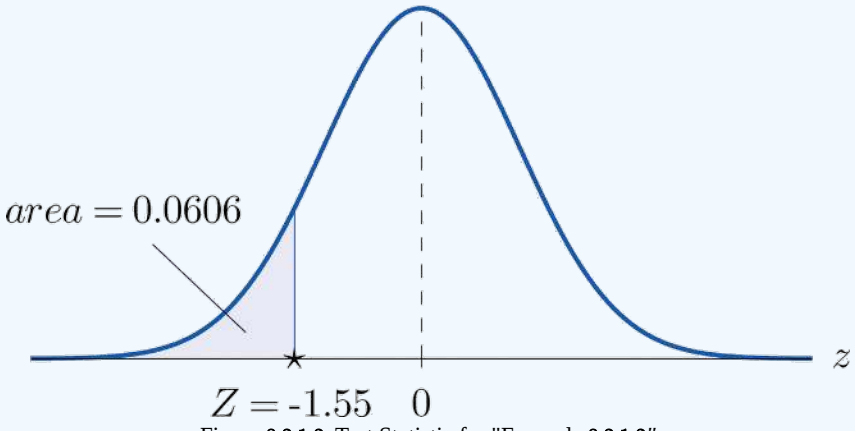


Figure 8.2.1.2: Test Statistic for "Example 8.2.1.2"

Example 8.2.1.3

Mr. Prospero has been teaching Algebra II from a particular textbook at Remote Isle High School for many years. Over the years students in his Algebra II classes have consistently scored an average of 67 on the end of course exam (EOC). This year Mr. Prospero used a new textbook in the hope that the average score on the EOC test would be higher. The average EOC test score of the 64 students who took Algebra II from Mr. Prospero this year had mean 69.4 and sample standard deviation 6.1. Determine whether these data provide sufficient evidence, at the 1% level of significance, to conclude that the average EOC test score is higher with the new textbook.

Solution

- **Step 1**. Let μ be the true average score on the EOC exam of all Mr. Prospero's students who take the Algebra II course with the new textbook. The natural statement that would be assumed true unless there were strong evidence to the contrary is that the new book is about the same as the old one. The alternative, which it takes evidence to establish, is that the new book is better, which corresponds to a higher value of μ. Thus the relevant test is

$$H_0 : \mu = 67$$
$$\text{vs}$$
$$H_a : \mu > 67 \ @ \ \alpha = 0.01$$

- **Step 2**. The sample is large and the population standard deviation is unknown. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the standard normal distribution.

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{69.4 - 67}{6.1/\sqrt{64}} = 3.15$$

- **Step 4**. The area of the right tail cut off by $z = 3.15$ is, by Figure 7.1.5, $1 - 0.9992 = 0.0008$, as shown in Figure 8.2.1.3. Since the test is right-tailed, the p-value is just this number, $p = 0.0008$.
- **Step 5**. Since $p = 0.0008 < 0.01 = \alpha$, the decision is to reject H_0. In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 1% level of significance, to conclude that the average EOC exam score of students taking the Algebra II course from Mr. Prospero using the new book is higher than the average score of those taking the course from him but using the old book.

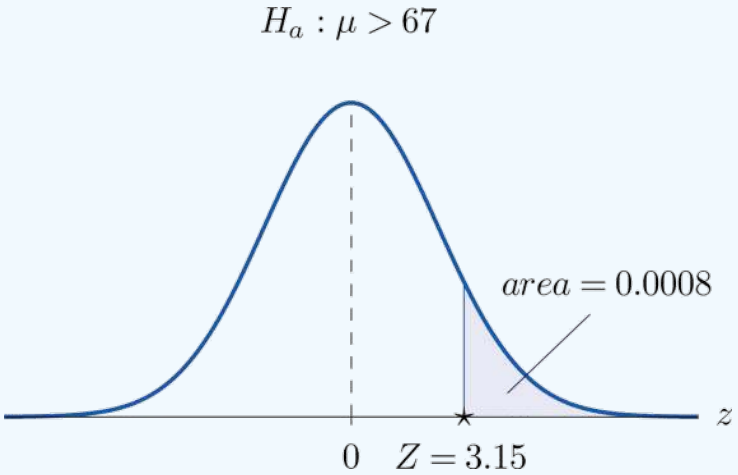


Figure 8.2.1.3: Test Statistic for "Example 8.2.1.3"

Example 8.2.1.4

For the surface water in a particular lake, local environmental scientists would like to maintain an average pH level at 7.4. Water samples are routinely collected to monitor the average pH level. If there is evidence of a shift in pH value, in either direction, then remedial action will be taken. On a particular day 30 water samples are taken and yield average pH reading of 7.7 with sample standard deviation 0.5. Determine, at the 1% level of significance, whether there is sufficient evidence in the sample to indicate that remedial action should be taken.

Solution

- **Step 1**. Let μ be the true average pH level at the time the samples were taken. The relevant test is

$$H_0 : \mu = 7.4$$
$$\text{vs}$$
$$H_a : \mu \neq 7.4 \ @ \ \alpha = 0.01$$

- **Step 2**. The sample is large and the population standard deviation is unknown. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the standard normal distribution.

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.7 - 7.4}{0.5/\sqrt{30}} = 3.29$$

- **Step 4**. The area of the right tail cut off by $z = 3.29$ is, by Figure 7.1.5, $1 - 0.9995 = 0.0005$, as illustrated in Figure 8.2.1.4. Since the test is two-tailed, the p-value is the double of this number, p=2×0.0005=0.0010.
- **Step 5**. Since $p = 0.0010 < 0.01 = \alpha$, the decision is to reject H_0. In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 1% level of significance, to conclude that the average pH of surface water in the lake is different from 7.4. That is, remedial action is indicated.

$$H_a : \mu \neq 7.4$$

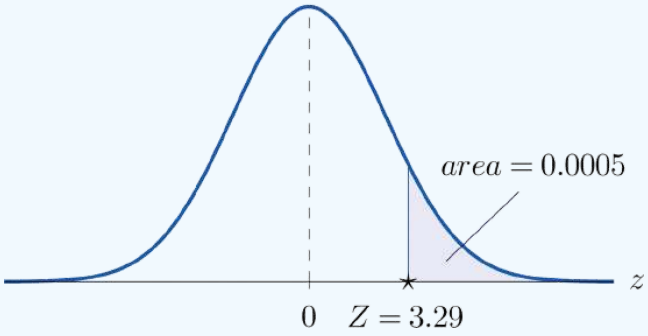


Figure 8.2.1.4:*Test Statistic for "Example* 8.2.1.4*"*

Key Takeaway

- The observed significance or p-value of a test is a measure of how inconsistent the sample result is with H_0 and in favor of H_a.
- The p-value approach to hypothesis testing means that one merely compares the p-value to α instead of constructing a rejection region.
- There is a systematic five-step procedure for the p-value approach to hypothesis testing.

This page titled 8.2.1: The Observed Significance of a Test is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **8.3: The Observed Significance of a Test** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

8.2.2: Small Sample Tests for a Population Mean

Learning Objectives

- To learn how to apply the five-step test procedure for test of hypotheses concerning a population mean when the sample size is small.

In the previous section hypotheses testing for population means was described in the case of large samples. The statistical validity of the tests was insured by the Central Limit Theorem, with essentially no assumptions on the distribution of the population. When sample sizes are small, as is often the case in practice, the Central Limit Theorem does not apply. One must then impose stricter assumptions on the population to give statistical validity to the test procedure. One common assumption is that the population from which the sample is taken has a normal probability distribution to begin with. Under such circumstances, if the population standard deviation is known, then the test statistic

$$\frac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}}$$

still has the standard normal distribution, as in the previous two sections. If σ is unknown and is approximated by the sample standard deviation s, then the resulting test statistic

$$\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$$

follows Student's t-distribution with $n-1$ degrees of freedom.

Standardized Test Statistics for Small Sample Hypothesis Tests Concerning a Single Population Mean

If σ is known:

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

If σ is unknown:

$$T=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

- The first test statistic (σ known) has the standard normal distribution.
- The second test statistic (σ unknown) has Student's t-distribution with $n-1$ degrees of freedom.
- The population must be normally distributed.

The distribution of the second standardized test statistic (the one containing s) and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure 8.2.2.1. This is just like Figure 8.2.1 except that now the critical values are from the t-distribution. Figure 8.2.1 still applies to the first standardized test statistic (the one containing (σ) since it follows the standard normal distribution.

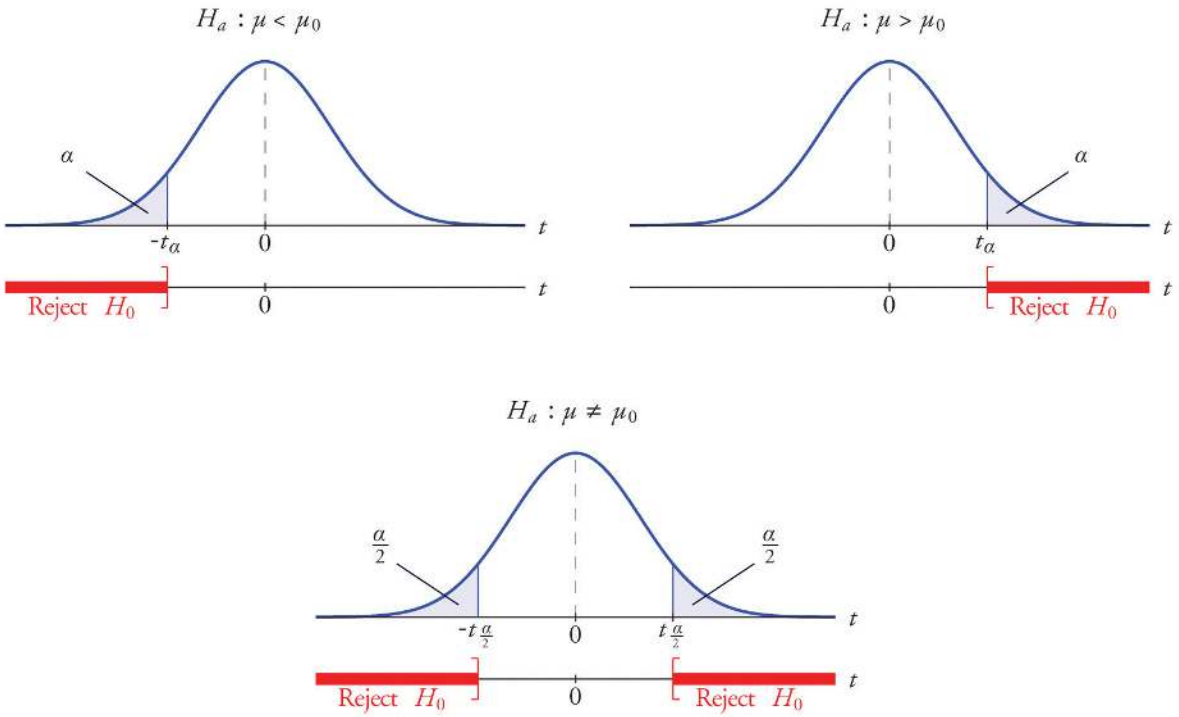


Figure **8.2.2.1**: Distribution of the Standardized Test Statistic and the Rejection Region

The p-value of a test of hypotheses for which the test statistic has Student's t-distribution can be computed using statistical software, but it is impractical to do so using tables, since that would require **30** tables analogous to Figure 7.1.5, one for each degree of freedom from **1** to **30**. Figure 7.1.6 can be used to approximate the p-value of such a test, and this is typically adequate for making a decision using the p-value approach to hypothesis testing, although not always. For this reason the tests in the two examples in this section will be made following the critical value approach to hypothesis testing summarized at the end of Section 8.1, but after each one we will show how the p-value approach could have been used.

Example 8.2.2.1

The price of a popular tennis racket at a national chain store is **$179**. Portia bought five of the same racket at an online auction site for the following prices:

155 179 175 175 161

Assuming that the auction prices of rackets are normally distributed, determine whether there is sufficient evidence in the sample, at the **5%** level of significance, to conclude that the average price of the racket is less than **$179** if purchased at an online auction.

Solution

- **Step 1**. The assertion for which evidence must be provided is that the average online price μ is less than the average price in retail stores, so the hypothesis test is

$$H_0 : \mu = 179$$
$$\text{vs}$$
$$H_a : \mu < 179 \text{ @ } \alpha = 0.05$$

- **Step 2**. The sample is small and the population standard deviation is unknown. Thus the test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  and has the Student t-distribution with $n - 1 = 5 - 1 = 4$ degrees of freedom.
- **Step 3**. From the data we compute $\bar{x} = 169$ and $s = 10.39$. Inserting these values into the formula for the test statistic gives

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{169 - 179}{10.39/\sqrt{5}} = -2.152$$

- **Step 4**. Since the symbol in H_a is "<" this is a left-tailed test, so there is a single critical value, $-t_\alpha = -t_{0.05}[df = 4]$. Reading from the row labeled $df = 4$ in Figure 7.1.6 its value is -2.132. The rejection region is $(-\infty, -2.132]$.
- **Step 5**. As shown in Figure 8.2.2.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the average price of such rackets purchased at online auctions is less than \$179.

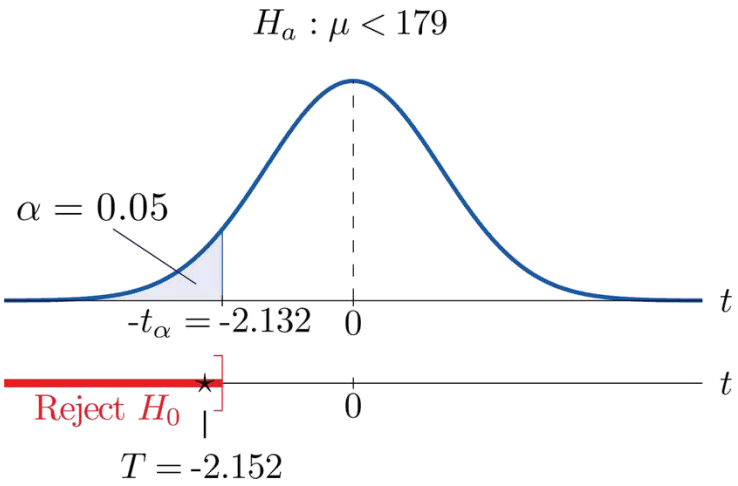


Figure 8.2.2.2: *Rejection Region and Test Statistic for "Example 8.2.2.1"*

To perform the test in Example 8.2.2.1 using the p-value approach, look in the row in Figure 7.1.6 with the heading $df = 4$ and search for the two t-values that bracket the unsigned value 2.152 of the test statistic. They are 2.132 and 2.776, in the columns with headings $t_{0.050}$ and $t_{0.025}$. They cut off right tails of area 0.050 and 0.025, so because 2.152 is between them it must cut off a tail of area between 0.050 and 0.025. By symmetry -2.152 cuts off a left tail of area between 0.050 and 0.025, hence the p-value corresponding to $t = -2.152$ is between 0.025 and 0.05. Although its precise value is unknown, it must be less than $\alpha = 0.05$, so the decision is to reject H_0.

Example 8.2.2.2

A small component in an electronic device has two small holes where another tiny part is fitted. In the manufacturing process the average distance between the two holes must be tightly controlled at 0.02 mm, else many units would be defective and wasted. Many times throughout the day quality control engineers take a small sample of the components from the production line, measure the distance between the two holes, and make adjustments if needed. Suppose at one time four units are taken and the distances are measured as

0.021 0.019 0.023 0.020

Determine, at the 1% level of significance, if there is sufficient evidence in the sample to conclude that an adjustment is needed. Assume the distances of interest are normally distributed.

Solution

- **Step 1**. The assumption is that the process is under control unless there is strong evidence to the contrary. Since a deviation of the average distance to either side is undesirable, the relevant test is

$$H_0 : \mu = 0.02$$
$$\text{vs}$$
$$H_a : \mu \neq 0.02 \ @ \ \alpha = 0.01$$

  where μ denotes the mean distance between the holes.
- **Step 2**. The sample is small and the population standard deviation is unknown. Thus the test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the Student t-distribution with $n - 1 = 4 - 1 = 3$ degrees of freedom.

- **Step 3**. From the data we compute $\bar{x} = 0.02075$ and $s = 0.00171$. Inserting these values into the formula for the test statistic gives

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.02075 - 0.02}{0.00171/\sqrt{4}} = 0.877$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm t_{\alpha/2} = -t_{0.005}[df = 3]$. Reading from the row in Figure 7.1.6 labeled $df = 3$ their values are ± 5.841. The rejection region is $(-\infty, -5.841] \cup [5.841, \infty)$.
- **Step 5**. As shown in Figure 8.2.2.3 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the mean distance between the holes in the component differs from 0.02 mm.

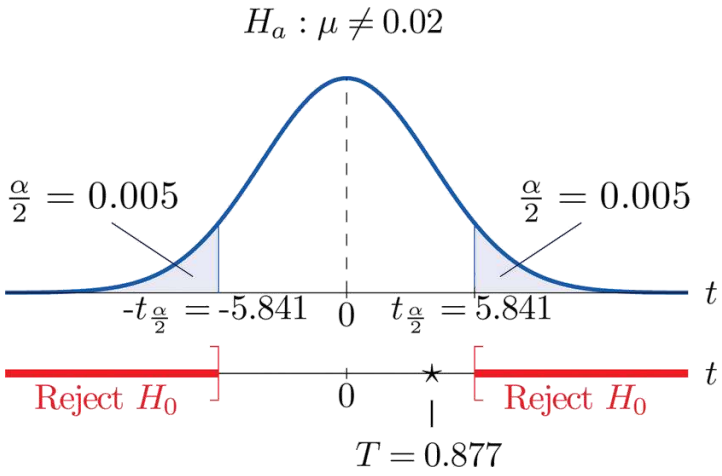


Figure 8.2.2.3: Rejection Region and Test Statistic for "Example 8.2.2.2"

To perform the test in "Example 8.2.2.2" using the p-value approach, look in the row in Figure 7.1.6 with the heading $df = 3$ and search for the two t-values that bracket the value 0.877 of the test statistic. Actually 0.877 is smaller than the smallest number in the row, which is 0.978, in the column with heading $t_{0.200}$. The value 0.978 cuts off a right tail of area 0.200, so because 0.877 is to its left it must cut off a tail of area greater than 0.200. Thus the p-value, which is the double of the area cut off (since the test is two-tailed), is greater than 0.400. Although its precise value is unknown, it must be greater than $\alpha = 0.01$, so the decision is not to reject H_0.

Key Takeaway

- There are two formulas for the test statistic in testing hypotheses about a population mean with small samples. One test statistic follows the standard normal distribution, the other Student's t-distribution.
- The population standard deviation is used if it is known, otherwise the sample standard deviation is used.
- Either five-step procedure, critical value or p-value approach, is used with either test statistic.

This page titled 8.2.2: Small Sample Tests for a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **8.4: Small Sample Tests for a Population Mean** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

8.3: Tests for a Population Proportion

Learning Objectives

- To learn how to apply the five-step critical value test procedure for test of hypotheses concerning a population proportion.
- To learn how to apply the five-step p-value test procedure for test of hypotheses concerning a population proportion.

Both the critical value approach and the p-value approach can be applied to test hypotheses about a population proportion p. The null hypothesis will have the form $H_0 : p = p_0$ for some specific number p_0 between 0 and 1. The alternative hypothesis will be one of the three inequalities

- $p < p_0$,
- $p > p_0$, or
- $p \neq p_0$

for the same number p_0 that appears in the null hypothesis.

The information in Section 6.3 gives the following formula for the test statistic and its distribution. In the formula p_0 is the numerical value of p that appears in the two hypotheses, $q_0 = 1 - p_0, \hat{p}$ is the sample proportion, and n is the sample size. Remember that the condition that the sample be large is not that n be at least 30 but that the interval

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

lie wholly within the interval $[0, 1]$.

Standardized Test Statistic for Large Sample Hypothesis Tests Concerning a Single Population Proportion

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_o}{n}}} \tag{8.3.1}$$

The test statistic has the standard normal distribution.

The distribution of the standardized test statistic and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure 8.3.1.

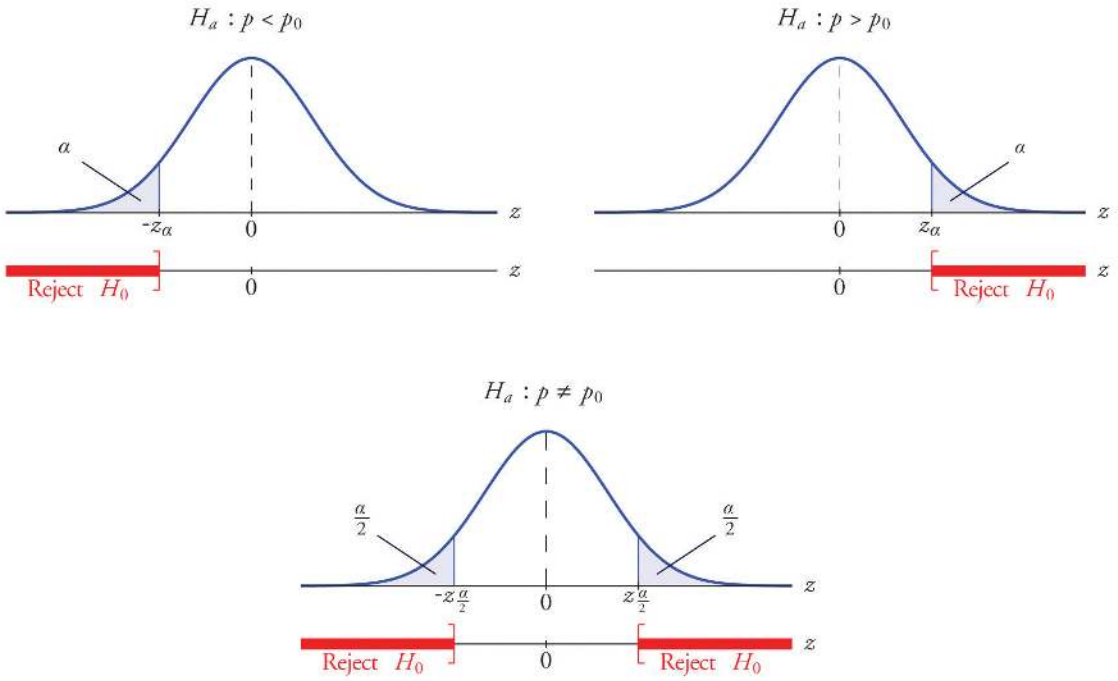


Figure 8.3.1: Distribution of the Standardized Test Statistic and the Rejection Region

Example 8.3.1

A soft drink maker claims that a majority of adults prefer its leading beverage over that of its main competitor's. To test this claim 500 randomly selected people were given the two beverages in random order to taste. Among them, 270 preferred the soft drink maker's brand, 211 preferred the competitor's brand, and 19 could not make up their minds. Determine whether there is sufficient evidence, at the 5% level of significance, to support the soft drink maker's claim against the default that the population is evenly split in its preference.

Solution

We will use the critical value approach to perform the test. The same test will be performed using the p-value approach in Example 8.3.3.

We must check that the sample is sufficiently large to validly perform the test. Since $\hat{p} = 270/500 = 0.54$,

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{500}} \approx 0.02$$

hence

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{8.3.2}$$

$$= [0.54 - (3)(0.02), 0.54 + (3)(0.02)] \tag{8.3.3}$$

$$= [0.48, 0.60] \subset [0, 1] \tag{8.3.4}$$

so the sample is sufficiently large.

- **Step 1.** The relevant test is

$$H_0 : p = 0.50$$

$$vs.$$

$$H_a : p > 0.50 \,@\, \alpha = 0.05$$

where p denotes the proportion of all adults who prefer the company's beverage over that of its competitor's beverage.

- **Step 2.** The test statistic (Equation 8.3.1) is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

and has the standard normal distribution.

- **Step 3.** The value of the test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \tag{8.3.5}$$

$$= \frac{0.54 - 0.50}{\sqrt{\frac{(0.50)(0.50)}{500}}} \tag{8.3.6}$$

$$= 1.789 \tag{8.3.7}$$

- **Step 4.** Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $z_\alpha = z_{0.05}$. Reading from the last line in Figure 7.1.6 its value is 1.645. The rejection region is $[1.645, \infty)$.
- **Step 5.** As shown in Figure 8.3.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that a majority of adults prefer the company's beverage to that of their competitor's.

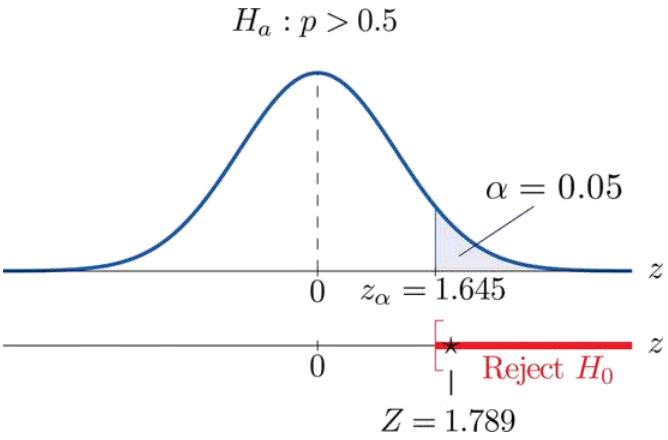


Figure 8.3.2: Rejection Region and Test Statistic for Example 8.3.1

Example 8.3.2

Globally the long-term proportion of newborns who are male is 51.46%. A researcher believes that the proportion of boys at birth changes under severe economic conditions. To test this belief randomly selected birth records of $5,000$ babies born during a period of economic recession were examined. It was found in the sample that 52.55% of the newborns were boys. Determine whether there is sufficient evidence, at the 10% level of significance, to support the researcher's belief.

Solution

We will use the critical value approach to perform the test. The same test will be performed using the p-value approach in Example 8.3.1.

The sample is sufficiently large to validly perform the test since

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.5255)(0.4745)}{5000}} \approx 0.01$$

hence

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{8.3.8}$$

$$= [0.5255 - 0.03, 0.5255 + 0.03] \tag{8.3.9}$$

$$= [0.4955, 0.5555] \subset [0, 1] \tag{8.3.10}$$

- **Step 1**. Let p be the true proportion of boys among all newborns during the recession period. The burden of proof is to show that severe economic conditions change it from the historic long-term value of 0.5146 rather than to show that it stays the same, so the hypothesis test is

$$H_0 : p = 0.5146$$

$$vs.$$

$$H_a : p \neq 0.5146 \,@\, \alpha = 0.10$$

- **Step 2.** The test statistic (Equation 8.3.1) is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

and has the standard normal distribution.

- **Step 3.** The value of the test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \tag{8.3.11}$$

$$= \frac{0.5255 - 0.5146}{\sqrt{\frac{(0.5146)(0.4854)}{5000}}} \tag{8.3.12}$$

$$= 1.542 \tag{8.3.13}$$

- **Step 4.** Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are a pair of critical values, $\pm z_{\alpha/2} = \pm z_{0.05} = \pm 1.645$. The rejection region is $(-\infty, -1.645] \cup [1.645, \infty)$.
- **Step 5.** As shown in Figure 8.3.3 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

  The data do not provide sufficient evidence, at the 10% level of significance, to conclude that the proportion of newborns who are male differs from the historic proportion in times of economic recession.

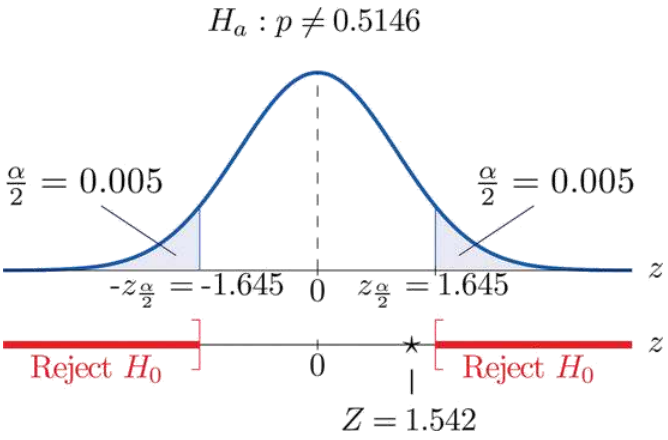


Figure 8.3.3: Rejection Region and Test Statistic for Example 8.3.2.

Example 8.3.3

Perform the test of Example 8.3.1 using the p-value approach.

Solution

We already know that the sample size is sufficiently large to validly perform the test.

- **Steps 1–3** of the five-step procedure described in Section 8.3 have already been done in Example 8.3.1 so we will not repeat them here, but only say that we know that the test is right-tailed and that value of the test statistic is $Z = 1.789$.
- **Step 4.** Since the test is right-tailed the p-value is the area under the standard normal curve cut off by the observed test statistic, $Z = 1.789$, as illustrated in Figure 8.3.4. By Figure 7.1.5 that area and therefore the p-value is $1 - 0.9633 = 0.0367$.
- **Step 5.** Since the p-value is less than $\alpha = 0.05$ the decision is to reject H_0.

$$H_a : p > 0.5$$

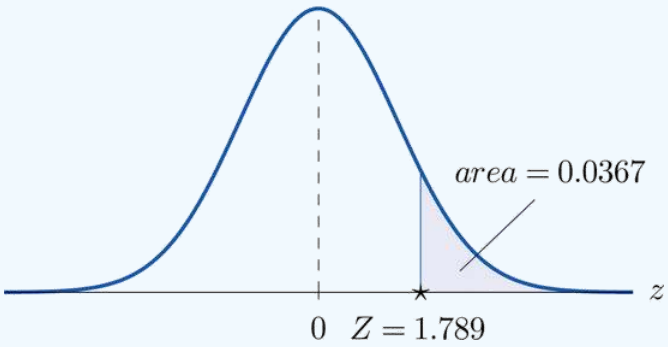


Figure 8.3.4: P-Value for Example 8.3.3

Example 8.3.4

Perform the test of Example 8.3.2 using the p-value approach.

Solution

We already know that the sample size is sufficiently large to validly perform the test.

- **Steps 1–3** of the five-step procedure described in Section 8.3 have already been done in Example 8.3.2. They tell us that the test is two-tailed and that value of the test statistic is $Z = 1.542$.
- **Step 4.** Since the test is two-tailed the p-value is the double of the area under the standard normal curve cut off by the observed test statistic, $Z = 1.542$. By Figure 7.1.5 that area is $1 - 0.9382 = 0.0618$, as illustrated in Figure 8.3.5, hence the p-value is $2 \times 0.0618 = 0.1236$.
- **Step 5.** Since the p-value is greater than $\alpha = 0.10$ the decision is not to reject H_0.

$$H_a : p \neq 0.5146$$

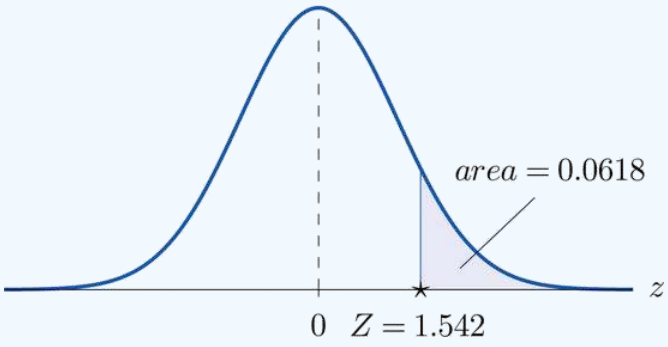


Figure 8.3.5: P-Value for Example 8.3.4

Key Takeaway

- There is one formula for the test statistic in testing hypotheses about a population proportion. The test statistic follows the standard normal distribution.
- Either five-step procedure, critical value or p-value approach, can be used.

This page titled 8.3: Tests for a Population Proportion is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **8.5: Large Sample Tests for a Population Proportion** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

8.E: Testing Hypotheses (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

8.1: The Elements of Hypothesis Testing

Q8.1.1

State the null and alternative hypotheses for each of the following situations. (That is, identify the correct number μ_0 and write $H_0 : \mu = \mu_0$ and the appropriate analogous expression for H_a.)

a. The average July temperature in a region historically has been $74.5^\circ F$. Perhaps it is higher now.
b. The average weight of a female airline passenger with luggage was 145 pounds ten years ago. The FAA believes it to be higher now.
c. The average stipend for doctoral students in a particular discipline at a state university is $\$14,756$. The department chairman believes that the national average is higher.
d. The average room rate in hotels in a certain region is $\$82.53$. A travel agent believes that the average in a particular resort area is different.
e. The average farm size in a predominately rural state was 69.4 acres. The secretary of agriculture of that state asserts that it is less today.

Q1.1.2

State the null and alternative hypotheses for each of the following situations. (That is, identify the correct number μ_0 and write $H_0 : \mu = \mu_0$ and the appropriate analogous expression for H_a.)

a. The average time workers spent commuting to work in Verona five years ago was 38.2 minutes. The Verona Chamber of Commerce asserts that the average is less now.
b. The mean salary for all men in a certain profession is $\$58,291$. A special interest group thinks that the mean salary for women in the same profession is different.
c. The accepted figure for the caffeine content of an 8-ounce cup of coffee is 133 mg. A dietitian believes that the average for coffee served in a local restaurants is higher.
d. The average yield per acre for all types of corn in a recent year was 161.9 bushels. An economist believes that the average yield per acre is different this year.
e. An industry association asserts that the average age of all self-described fly fishermen is 42.8 years. A sociologist suspects that it is higher.

Q1.1.3

Describe the two types of errors that can be made in a test of hypotheses.

Q1.1.4

Under what circumstance is a test of hypotheses certain to yield a correct decision?

Answers

1. a. $H_0 : \mu = 74.5 \; vs \; H_a : \mu > 74.5$
   b. $H_0 : \mu = 145 \; vs \; H_a : \mu > 145$
   c. $H_0 : \mu = 14756 \; vs \; H_a : \mu > 14756$
   d. $H_0 : \mu = 82.53 \; vs \; H_a : \mu \neq 82.53$
   e. $H_0 : \mu = 69.4 \; vs \; H_a : \mu < 69.4$
2.
3. A Type I error is made when a true H_0 is rejected. A Type II error is made when a false H_0 is not rejected.

8.2: Large Sample Tests for a Population Mean

Basic

1. Find the rejection region (for the standardized test statistic) for each hypothesis test.
   a. $H_0 : \mu = 27$ *vs* $H_a : \mu < 27$ @ $\alpha = 0.05$
   b. $H_0 : \mu = 52$ *vs* $H_a : \mu \neq 52$ @ $\alpha = 0.05$
   c. $H_0 : \mu = -105$ *vs* $H_a : \mu > -105$ @ $\alpha = 0.10$
   d. $H_0 : \mu = 78.8$ *vs* $H_a : \mu \neq 78.8$ @ $\alpha = 0.10$
2. Find the rejection region (for the standardized test statistic) for each hypothesis test.
   a. $H_0 : \mu = 17$ *vs* $H_a : \mu < 17$ @ $\alpha = 0.01$
   b. $H_0 : \mu = 880$ *vs* $H_a : \mu \neq 880$ @ $\alpha = 0.01$
   c. $H_0 : \mu = -12$ *vs* $H_a : \mu > -12$ @ $\alpha = 0.05$
   d. $H_0 : \mu = 21.1$ *vs* $H_a : \mu \neq 21.1$ @ $\alpha = 0.05$
3. Find the rejection region (for the standardized test statistic) for each hypothesis test. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = 141$ *vs* $H_a : \mu < 141$ @ $\alpha = 0.20$
   b. $H_0 : \mu = -54$ *vs* $H_a : \mu < -54$ @ $\alpha = 0.05$
   c. $H_0 : \mu = 98.6$ *vs* $H_a : \mu \neq 98.6$ @ $\alpha = 0.05$
   d. $H_0 : \mu = 3.8$ *vs* $H_a : \mu > 3.8$ @ $\alpha = 0.001$
4. Find the rejection region (for the standardized test statistic) for each hypothesis test. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = -62$ *vs* $H_a : \mu \neq -62$ @ $\alpha = 0.005$
   b. $H_0 : \mu = 73$ *vs* $H_a : \mu > 73$ @ $\alpha = 0.001$
   c. $H_0 : \mu = 1124$ *vs* $H_a : \mu < 1124$ @ $\alpha = 0.001$
   d. $H_0 : \mu = 0.12$ *vs* $H_a : \mu \neq 0.12$ @ $\alpha = 0.001$
5. Compute the value of the test statistic for the indicated test, based on the information given.
   a. Testing $H_0 : \mu = 72.2$ *vs* $H_a : \mu > 72.2$, σ unknown $n = 55$, $\bar{x} = 75.1$, $s = 9.25$
   b. Testing $H_0 : \mu = 58$ *vs* $H_a : \mu > 58$, $\sigma = 1.22$ $n = 40$, $\bar{x} = 58.5$, $s = 1.29$
   c. Testing $H_0 : \mu = -19.5$ *vs* $H_a : \mu < -19.5$, σ unknown $n = 30$, $\bar{x} = -23.2$, $s = 9.55$
   d. Testing $H_0 : \mu = 805$ *vs* $H_a : \mu \neq 805$, $\sigma = 37.5$ $n = 75$, $\bar{x} = 818$, $s = 36.2$
6. Compute the value of the test statistic for the indicated test, based on the information given.
   a. Testing $H_0 : \mu = 342$ *vs* $H_a : \mu < 342$, $\sigma = 11.2$ $n = 40$, $\bar{x} = 339$, $s = 10.3$
   b. Testing $H_0 : \mu = 105$ *vs* $H_a : \mu > 105$, $\sigma = 5.3$ $n = 80$, $\bar{x} = 107$, $s = 5.1$
   c. Testing $H_0 : \mu = -13.5$ *vs* $H_a : \mu \neq -13.5$, σ unknown $n = 32$, $\bar{x} = -13.8$, $s = 1.5$
   d. Testing $H_0 : \mu = 28$ *vs* $H_a : \mu \neq 28$, σ unknown $n = 68$, $\bar{x} = 27.8$, $s = 1.3$
7. Perform the indicated test of hypotheses, based on the information given.
   a. Test $H_0 : \mu = 212$ *vs* $H_a : \mu < 212$ @ $\alpha = 0.10$, σ unknown $n = 36$, $\bar{x} = 211.2$, $s = 2.2$
   b. Test $H_0 : \mu = -18$ *vs* $H_a : \mu > -18$ @ $\alpha = 0.05$, $\sigma = 3.3$ $n = 44$, $\bar{x} = -17.2$, $s = 3.1$
   c. Test $H_0 : \mu = 24$ *vs* $H_a : \mu \neq 24$ @ $\alpha = 0.02$, σ unknown $n = 50$, $\bar{x} = 22.8$, $s = 1.9$
8. Perform the indicated test of hypotheses, based on the information given.
   a. Test $H_0 : \mu = 105$ *vs* $H_a : \mu > 105$ @ $\alpha = 0.05$, σ unknown $n = 30$, $\bar{x} = 108$, $s = 7.2$
   b. Test $H_0 : \mu = 21.6$ *vs* $H_a : \mu < 21.6$ @ $\alpha = 0.01$, σ unknown $n = 78$, $\bar{x} = 20.5$, $s = 3.9$
   c. Test $H_0 : \mu = -375$ *vs* $H_a : \mu \neq -375$ @ $\alpha = 0.01$, $\sigma = 18.5$ $n = 31$, $\bar{x} = -388$, $s = 18.0$

Applications

9. In the past the average length of an outgoing telephone call from a business office has been **143** seconds. A manager wishes to check whether that average has decreased after the introduction of policy changes. A sample of **100** telephone calls produced a mean of **133** seconds, with a standard deviation of **35** seconds. Perform the relevant test at the **1%** level of significance.

10. The government of an impoverished country reports the mean age at death among those who have survived to adulthood as 66.2 years. A relief agency examines 30 randomly selected deaths and obtains a mean of 62.3 years with standard deviation 8.1 years. Test whether the agency's data support the alternative hypothesis, at the 1% level of significance, that the population mean is less than 66.2.
11. The average household size in a certain region several years ago was 3.14 persons. A sociologist wishes to test, at the 5% level of significance, whether it is different now. Perform the test using the information collected by the sociologist: in a random sample of 75 households, the average size was 2.98 persons, with sample standard deviation 0.82 person.
12. The recommended daily calorie intake for teenage girls is $2,200$ calories/day. A nutritionist at a state university believes the average daily caloric intake of girls in that state to be lower. Test that hypothesis, at the 5% level of significance, against the null hypothesis that the population average is $2,200$ calories/day using the following sample data: $n = 36,\ \bar{x} = 2,150,\ s = 203$
13. An automobile manufacturer recommends oil change intervals of $3,000$ miles. To compare actual intervals to the recommendation, the company randomly samples records of 50 oil changes at service facilities and obtains sample mean $3,752$ miles with sample standard deviation 638 miles. Determine whether the data provide sufficient evidence, at the 5% level of significance, that the population mean interval between oil changes exceeds $3,000$ miles.
14. A medical laboratory claims that the mean turn-around time for performance of a battery of tests on blood samples is 1.88 business days. The manager of a large medical practice believes that the actual mean is larger. A random sample of 45 blood samples yielded mean 2.09 and sample standard deviation 0.13 day. Perform the relevant test at the 10% level of significance, using these data.
15. A grocery store chain has as one standard of service that the mean time customers wait in line to begin checking out not exceed 2 minutes. To verify the performance of a store the company measures the waiting time in 30 instances, obtaining mean time 2.17 minutes with standard deviation 0.46 minute. Use these data to test the null hypothesis that the mean waiting time is 2 minutes versus the alternative that it exceeds 2 minutes, at the 10% level of significance.
16. A magazine publisher tells potential advertisers that the mean household income of its regular readership is $\$61,500$. An advertising agency wishes to test this claim against the alternative that the mean is smaller. A sample of 40 randomly selected regular readers yields mean income $\$59,800$ with standard deviation $\$5,850$. Perform the relevant test at the 1% level of significance.
17. Authors of a computer algebra system wish to compare the speed of a new computational algorithm to the currently implemented algorithm. They apply the new algorithm to 50 standard problems; it averages 8.16 seconds with standard deviation 0.17 second. The current algorithm averages 8.21 seconds on such problems. Test, at the 1% level of significance, the alternative hypothesis that the new algorithm has a lower average time than the current algorithm.
18. A random sample of the starting salaries of 35 randomly selected graduates with bachelor's degrees last year gave sample mean and standard deviation $\$41,202$ and $\$7,621$, respectively. Test whether the data provide sufficient evidence, at the 5% level of significance, to conclude that the mean starting salary of all graduates last year is less than the mean of all graduates two years before, $\$43,589$.

Additional Exercises

19. The mean household income in a region served by a chain of clothing stores is $\$48,750$. In a sample of 40 customers taken at various stores the mean income of the customers was $\$51,505$ with standard deviation $\$6,852$.
    a. Test at the 10% level of significance the null hypothesis that the mean household income of customers of the chain is $\$48,750$ against that alternative that it is different from $\$48,750$.
    b. The sample mean is greater than $\$48,750$, suggesting that the actual mean of people who patronize this store is greater than $\$48,750$. Perform this test, also at the 10% level of significance. (The computation of the test statistic done in part (a) still applies here.)
20. The labor charge for repairs at an automobile service center are based on a standard time specified for each type of repair. The time specified for replacement of universal joint in a drive shaft is one hour. The manager reviews a sample of 30 such repairs. The average of the actual repair times is 0.86 hour with standard deviation 0.32 hour.
    a. Test at the 1% level of significance the null hypothesis that the actual mean time for this repair differs from one hour.
    b. The sample mean is less than one hour, suggesting that the mean actual time for this repair is less than one hour. Perform this test, also at the 1% level of significance. (The computation of the test statistic done in part (a) still applies here.)

Large Data Set Exercises

Large Data Set missing from the original

21. Large **Data Set 1** records the SAT scores of $1,000$ students. Regarding it as a random sample of all high school students, use it to test the hypothesis that the population mean exceeds $1,510$, at the 1% level of significance. (The null hypothesis is that $\mu = 1510$).
22. Large **Data Set 1** records the GPAs of $1,000$ college students. Regarding it as a random sample of all college students, use it to test the hypothesis that the population mean is less than 2.50, at the 10% level of significance. (The null hypothesis is that $\mu = 2.50$).
23. Large **Data Set 1** lists the SAT scores of $1,000$ students.
    a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.
    b. Regard the first 50 students in the data set as a random sample drawn from the population of part (a) and use it to test the hypothesis that the population mean exceeds $1,510$, at the 10% level of significance. (The null hypothesis is that $\mu = 1510$).
    c. Is your conclusion in part (b) in agreement with the true state of nature (which by part (a) you know), or is your decision in error? If your decision is in error, is it a Type I error or a Type II error?
24. Large **Data Set 1** lists the GPAs of $1,000$ students.
    a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.
    b. Regard the first 50 students in the data set as a random sample drawn from the population of part (a) and use it to test the hypothesis that the population mean is less than 2.50, at the 10% level of significance. (The null hypothesis is that $\mu = 2.50$).
    c. Is your conclusion in part (b) in agreement with the true state of nature (which by part (a) you know), or is your decision in error? If your decision is in error, is it a Type I error or a Type II error?

Answers

1. a. $Z \le -1.645$
   b. $Z \le -1.645$ *or* $Z \ge 1.96$
   c. $Z \ge 1.28$
   d. $Z \le -1.645$ *or* $Z \ge 1.645$
2.
3. a. $Z \le -0.84$
   b. $Z \le -1.645$
   c. $Z \le -1.96$ *or* $Z \ge 1.96$
   d. $Z \ge 3.1$
4.
5. a. $Z = 2.235$
   b. $Z = 2.592$
   c. $Z = -2.122$
   d. $Z = 3.002$
6.
7. a. $Z = -2.18,\ -z_{0.10} = -1.28,$ reject H_0
   b. $Z = 1.61,\ z_{0.05} = 1.645,$ do not reject H_0
   c. $Z = -4.47,\ -z_{0.01} = -2.33,$ reject H_0
8.
9. $Z = -2.86,\ -z_{0.01} = -2.33,$ reject H_0
10.
11. $Z = -1.69,\ -z_{0.025} = -1.96,$ do not reject H_0
12.
13. $Z = 8.33,\ z_{0.05} = 1.645,$ reject H_0

14.
15. $Z = 2.02,\ z_{0.10} = 1.28,$ reject H_0
16.
17. $Z = -2.08,\ -z_{0.01} = -2.33,$ do not reject H_0
18.
19. a. $Z = 2.54,\ z_{0.05} = 1.645,$ reject H_0
    b. $Z = 2.54,\ z_{0.10} = 1.28,$ reject H_0
20.
21. $H_0 : \mu = 1510$ *vs* $H_a : \mu > 1510$. Test Statistic: $Z = 2.7882$. Rejection Region: $[2.33, \infty)$. Decision: Reject H_0.
22.
23. a. $\mu_0 = 1528.74$
    b. $H_0 : \mu = 1510$ *vs* $H_a : \mu > 1510$. Test Statistic: $Z = -1.41$. Rejection Region: $[1.28, \infty)$. Decision: Fail to reject H_0.
    c. No, it is a Type II error.

8.3: The Observed Significance of a Test

Basic

1. Compute the observed significance of each test.
   a. Testing $H_0 : \mu = 54.7$ *vs* $H_a : \mu < 54.7,$ test statistic $z = -1.72$
   b. Testing $H_0 : \mu = 195$ *vs* $H_a : \mu \neq 195,$ test statistic $z = -2.07$
   c. Testing $H_0 : \mu = -45$ *vs* $H_a : \mu > -45,$ test statistic $z = 2.54$
2. Compute the observed significance of each test.
   a. Testing $H_0 : \mu = 0$ *vs* $H_a : \mu \neq 0,$ test statistic $z = 2.82$
   b. Testing $H_0 : \mu = 18.4$ *vs* $H_a : \mu < 18.4,$ test statistic $z = -1.74$
   c. Testing $H_0 : \mu = 63.85$ *vs* $H_a : \mu > 63.85,$ test statistic $z = 1.93$
3. Compute the observed significance of each test. (Some of the information given might not be needed.)
   a. Testing $H_0 : \mu = 27.5$ *vs* $H_a : \mu > 27.5,\ n = 49,\ \bar{x} = 28.9,\ s = 3.14,$ test statistic $z = 3.12$
   b. Testing $H_0 : \mu = 581$ *vs* $H_a : \mu < 581,\ n = 32,\ \bar{x} = 560,\ s = 47.8,$ test statistic $z = -2.49$
   c. Testing $H_0 : \mu = 138.5$ *vs* $H_a : \mu \neq 138.5,\ n = 44,\ \bar{x} = 137.6,\ s = 2.45,$ test statistic $z = -2.44$
4. Compute the observed significance of each test. (Some of the information given might not be needed.)
   a. Testing $H_0 : \mu = -17.9$ *vs* $H_a : \mu < -17.9,\ n = 34,\ \bar{x} = -18.2,\ s = 0.87,$ test statistic $z = -2.01$
   b. Testing $H_0 : \mu = 5.5$ *vs* $H_a : \mu \neq 5.5,\ n = 56,\ \bar{x} = 7.4,\ s = 4.82,$ test statistic $z = 2.95$
   c. Testing $H_0 : \mu = 1255$ *vs* $H_a : \mu > 1255,\ n = 152,\ \bar{x} = 1257,\ s = 7.5,$ test statistic $z = 3.29$
5. Make the decision in each test, based on the information provided.
   a. Testing $H_0 : \mu = 82.9$ *vs* $H_a : \mu < 82.9$ @ $\alpha = 0.05$, observed significance $p = 0.038$
   b. Testing $H_0 : \mu = 213.5$ *vs* $H_a : \mu \neq 213.5$ @ $\alpha = 0.01$, observed significance $p = 0.038$
6. Make the decision in each test, based on the information provided.
   a. Testing $H_0 : \mu = 31.4$ *vs* $H_a : \mu > 31.4$ @ $\alpha = 0.10$, observed significance $p = 0.062$
   b. Testing $H_0 : \mu = -75.5$ *vs* $H_a : \mu < -75.5$ @ $\alpha = 0.05$, observed significance $p = 0.062$

Applications

7. A lawyer believes that a certain judge imposes prison sentences for property crimes that are longer than the state average 11.7 months. He randomly selects 36 of the judge's sentences and obtains mean 13.8 and standard deviation 3.9 months.
   a. Perform the test at the 1% level of significance using the critical value approach.
   b. Compute the observed significance of the test.
   c. Perform the test at the 1% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).

8. In a recent year the fuel economy of all passenger vehicles was 19.8 mpg. A trade organization sampled 50 passenger vehicles for fuel economy and obtained a sample mean of 20.1 mpg with standard deviation 2.45 mpg. The sample mean 20.1 exceeds 19.8, but perhaps the increase is only a result of sampling error.
   a. Perform the relevant test of hypotheses at the 20% level of significance using the critical value approach.
   b. Compute the observed significance of the test.
   c. Perform the test at the 20% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
9. The mean score on a 25-point placement exam in mathematics used for the past two years at a large state university is 14.3. The placement coordinator wishes to test whether the mean score on a revised version of the exam differs from 14.3. She gives the revised exam to 30 entering freshmen early in the summer; the mean score is 14.6 with standard deviation 2.4.
   a. Perform the test at the 10% level of significance using the critical value approach.
   b. Compute the observed significance of the test.
   c. Perform the test at the 10% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
10. The mean increase in word family vocabulary among students in a one-year foreign language course is 576 word families. In order to estimate the effect of a new type of class scheduling, an instructor monitors the progress of 60 students; the sample mean increase in word family vocabulary of these students is 542 word families with sample standard deviation 18 word families.
    a. Test at the 5% level of significance whether the mean increase with the new class scheduling is different from 576 word families, using the critical value approach.
    b. Compute the observed significance of the test.
    c. Perform the test at the 5% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
11. The mean yield for hard red winter wheat in a certain state is 44.8 bu/acre. In a pilot program a modified growing scheme was introduced on 35 independent plots. The result was a sample mean yield of 45.4 bu/acre with sample standard deviation 1.6 bu/acre, an apparent increase in yield.
    a. Test at the 5% level of significance whether the mean yield under the new scheme is greater than 44.8 bu/acre, using the critical value approach.
    b. Compute the observed significance of the test.
    c. Perform the test at the 5% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
12. The average amount of time that visitors spent looking at a retail company's old home page on the world wide web was 23.6 seconds. The company commissions a new home page. On its first day in place the mean time spent at the new page by $7,628$ visitors was 23.5 seconds with standard deviation 5.1 seconds.
    a. Test at the 5% level of significance whether the mean visit time for the new page is less than the former mean of 23.6 seconds, using the critical value approach.
    b. Compute the observed significance of the test.
    c. Perform the test at the 5% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).

Answers

1. a. p-value $= 0.0427$
   b. p-value $= 0.0384$
   c. p-value $= 0.0055$
2.
3. a. p-value $= 0.0009$
   b. p-value $= 0.0064$
   c. p-value $= 0.0146$

4.

5. a. reject H_0
   b. do not reject H_0

6.

7. a. $Z = 3.23,\ z_{0.01} = 2.33$, reject H_0
   b. p-value $= 0.0006$
   c. reject H_0

8.

9. a. $Z = 0.68,\ z_{0.05} = 1.645$, do not reject H_0
   b. p-value $= 0.4966$
   c. do not reject H_0

10.

11. a. $Z = 2.22,\ z_{0.05} = 1.645$, reject H_0
    b. p-value $= 0.0132$
    c. reject H_0

8.4: Small Sample Tests for a Population Mean

Basic

1. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed.
   a. $H_0 : \mu = 27\ vs\ H_a : \mu < 27\ @\ \alpha = 0.05,\ n = 12,\ \sigma = 2.2$
   b. $H_0 : \mu = 52\ vs\ H_a : \mu \neq 52\ @\ \alpha = 0.05,\ n = 6,\ \sigma$ unknown
   c. $H_0 : \mu = -105\ vs\ H_a : \mu > -105\ @\ \alpha = 0.10,\ n = 24,\ \sigma$ unknown
   d. $H_0 : \mu = 78.8\ vs\ H_a : \mu \neq 78.8\ @\ \alpha = 0.10,\ n = 8,\ \sigma = 1.7$
2. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed.
   a. $H_0 : \mu = 17\ vs\ H_a : \mu < 17\ @\ \alpha = 0.01,\ n = 26,\ \sigma = 0.94$
   b. $H_0 : \mu = 880\ vs\ H_a : \mu \neq 880\ @\ \alpha = 0.01,\ n = 4,\ \sigma$ unknown
   c. $H_0 : \mu = -12\ vs\ H_a : \mu > -12\ @\ \alpha = 0.05,\ n = 18,\ \sigma = 1.1$
   d. $H_0 : \mu = 21.1\ vs\ H_a : \mu \neq 21.1\ @\ \alpha = 0.05,\ n = 23,\ \sigma$ unknown
3. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = 141\ vs\ H_a : \mu < 141\ @\ \alpha = 0.20,\ n = 29,\ \sigma$ unknown
   b. $H_0 : \mu = -54\ vs\ H_a : \mu < -54\ @\ \alpha = 0.05,\ n = 15,\ \sigma = 1.9$
   c. $H_0 : \mu = 98.6\ vs\ H_a : \mu \neq 98.6\ @\ \alpha = 0.05,\ n = 12,\ \sigma$ unknown
   d. $H_0 : \mu = 3.8\ vs\ H_a : \mu > 3.8\ @\ \alpha = 0.001,\ n = 27,\ \sigma$ unknown
4. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = -62\ vs\ H_a : \mu \neq -62\ @\ \alpha = 0.005,\ n = 8,\ \sigma$ unknown
   b. $H_0 : \mu = 73\ vs\ H_a : \mu > 73\ @\ \alpha = 0.001,\ n = 22,\ \sigma$ unknown
   c. $H_0 : \mu = 1124\ vs\ H_a : \mu < 1124\ @\ \alpha = 0.001,\ n = 21,\ \sigma$ unknown
   d. $H_0 : \mu = 0.12\ vs\ H_a : \mu \neq 0.12\ @\ \alpha = 0.001,\ n = 14,\ \sigma = 0.026$
5. A random sample of size 20 drawn from a normal population yielded the following results: $\bar{x} = 49.2,\ s = 1.33$
   a. Test $H_0 : \mu = 50\ vs\ H_a : \mu \neq 50\ @\ \alpha = 0.01$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.

6. A random sample of size 16 drawn from a normal population yielded the following results: $\bar{x} = -0.96,\ s = 1.07$
   a. Test $H_0 : \mu = 0\ vs\ H_a : \mu < 0\ @\ \alpha = 0.001$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.
7. A random sample of size 8 drawn from a normal population yielded the following results: $\bar{x} = 289,\ s = 46$
   a. Test $H_0 : \mu = 250\ vs\ H_a : \mu > 250\ @\ \alpha = 0.05$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.
8. A random sample of size 12 drawn from a normal population yielded the following results: $\bar{x} = 86.2,\ s = 0.63$
   a. Test $H_0 : \mu = 85.5\ vs\ H_a : \mu \neq 85.5\ @\ \alpha = 0.01$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.

Applications

9. Researchers wish to test the efficacy of a program intended to reduce the length of labor in childbirth. The accepted mean labor time in the birth of a first child is 15.3 hours. The mean length of the labors of 13 first-time mothers in a pilot program was 8.8 hours with standard deviation 3.1 hours. Assuming a normal distribution of times of labor, test at the 10% level of significance test whether the mean labor time for all women following this program is less than 15.3 hours.
10. A dairy farm uses the somatic cell count (SCC) report on the milk it provides to a processor as one way to monitor the health of its herd. The mean SCC from five samples of raw milk was $250,000$ cells per milliliter with standard deviation $37,500$ cell/ml. Test whether these data provide sufficient evidence, at the 10% level of significance, to conclude that the mean SCC of all milk produced at the dairy exceeds that in the previous report, $210,250$ cell/ml. Assume a normal distribution of SCC.
11. Six coins of the same type are discovered at an archaeological site. If their weights on average are significantly different from 5.25 grams then it can be assumed that their provenance is not the site itself. The coins are weighed and have mean 4.73 g with sample standard deviation 0.18 g. Perform the relevant test at the 0.1% ($1/10$th of 1%) level of significance, assuming a normal distribution of weights of all such coins.
12. An economist wishes to determine whether people are driving less than in the past. In one region of the country the number of miles driven per household per year in the past was 18.59 thousand miles. A sample of 15 households produced a sample mean of 16.23 thousand miles for the last year, with sample standard deviation 4.06 thousand miles. Assuming a normal distribution of household driving distances per year, perform the relevant test at the 5% level of significance.
13. The recommended daily allowance of iron for females aged $19 - 50$ is 18 mg/day. A careful measurement of the daily iron intake of 15 women yielded a mean daily intake of 16.2 mg with sample standard deviation 4.7 mg.
    a. Assuming that daily iron intake in women is normally distributed, perform the test that the actual mean daily intake for all women is different from 18 mg/day, at the 10% level of significance.
    b. The sample mean is less than 18, suggesting that the actual population mean is less than 18 mg/day. Perform this test, also at the 10% level of significance. (The computation of the test statistic done in part (a) still applies here.)
14. The target temperature for a hot beverage the moment it is dispensed from a vending machine is $170^\circ F$. A sample of ten randomly selected servings from a new machine undergoing a pre-shipment inspection gave mean temperature $173^\circ F$ with sample standard deviation $6.3^\circ F$.
    a. Assuming that temperature is normally distributed, perform the test that the mean temperature of dispensed beverages is different from $170^\circ F$, at the 10% level of significance.
    b. The sample mean is greater than 170, suggesting that the actual population mean is greater than $170^\circ F$. Perform this test, also at the 10% level of significance. (The computation of the test statistic done in part (a) still applies here.)
15. The average number of days to complete recovery from a particular type of knee operation is 123.7 days. From his experience a physician suspects that use of a topical pain medication might be lengthening the recovery time. He randomly selects the records of seven knee surgery patients who used the topical medication. The times to total recovery were:

$$128 \quad 135 \quad 121 \quad 142 \quad 126 \quad 151 \quad 123 \tag{8.E.1}$$

    a. Assuming a normal distribution of recovery times, perform the relevant test of hypotheses at the 10% level of significance.

b. Would the decision be the same at the 5% level of significance? Answer either by constructing a new rejection region (critical value approach) or by estimating the p-value of the test in part (a) and comparing it to α.

16. A 24-hour advance prediction of a day's high temperature is "unbiased" if the long-term average of the error in prediction (true high temperature minus predicted high temperature) is zero. The errors in predictions made by one meteorological station for 20 randomly selected days were:

$$\begin{array}{ccccc} 2 & 0 & -3 & 1 & -2 \\ 1 & 0 & -1 & 1 & -1 \\ -4 & 1 & 1 & -4 & 0 \\ -4 & -3 & -4 & 2 & 2 \end{array} \tag{8.E.2}$$

a. Assuming a normal distribution of errors, test the null hypothesis that the predictions are unbiased (the mean of the population of all errors is 0) versus the alternative that it is biased (the population mean is not 0), at the 1% level of significance.

b. Would the decision be the same at the 5% level of significance? The 10% level of significance? Answer either by constructing new rejection regions (critical value approach) or by estimating the p-value of the test in part (a) and comparing it to α.

17. Pasteurized milk may not have a standardized plate count (SPC) above $20,000$ colony-forming bacteria per milliliter (cfu/ml). The mean SPC for five samples was $21,500$ cfu/ml with sample standard deviation 750 cfu/ml. Test the null hypothesis that the mean SPC for this milk is $20,000$ versus the alternative that it is greater than $20,000$, at the 10% level of significance. Assume that the SPC follows a normal distribution.

18. One water quality standard for water that is discharged into a particular type of stream or pond is that the average daily water temperature be at most $18°F$. Six samples taken throughout the day gave the data:

$$16.8 \quad 21.5 \quad 19.1 \quad 12.8 \quad 18.0 \quad 20.7 \tag{8.E.3}$$

The sample mean exceeds $\bar{x} = 18.15$, but perhaps this is only sampling error. Determine whether the data provide sufficient evidence, at the 10% level of significance, to conclude that the mean temperature for the entire day exceeds $18°F$.

Additional Exercises

19. A calculator has a built-in algorithm for generating a random number according to the standard normal distribution. Twenty-five numbers thus generated have mean 0.15 and sample standard deviation 0.94. Test the null hypothesis that the mean of all numbers so generated is 0 versus the alternative that it is different from 0, at the 20% level of significance. Assume that the numbers do follow a normal distribution.

20. At every setting a high-speed packing machine delivers a product in amounts that vary from container to container with a normal distribution of standard deviation 0.12 ounce. To compare the amount delivered at the current setting to the desired amount 64.1 ounce, a quality inspector randomly selects five containers and measures the contents of each, obtaining sample mean 63.9 ounces and sample standard deviation 0.10 ounce. Test whether the data provide sufficient evidence, at the 5% level of significance, to conclude that the mean of all containers at the current setting is less than 64.1 ounces.

21. A manufacturing company receives a shipment of $1,000$ bolts of nominal shear strength $4,350$ lb. A quality control inspector selects five bolts at random and measures the shear strength of each. The data are:

$$4,320 \quad 4,290 \quad 4,360 \quad 4,350 \quad 4,320 \tag{8.E.4}$$

a. Assuming a normal distribution of shear strengths, test the null hypothesis that the mean shear strength of all bolts in the shipment is $4,350$ lb versus the alternative that it is less than $4,350$ lb, at the 10% level of significance.

b. Estimate the p-value (observed significance) of the test of part (a).

c. Compare the p-value found in part (b) to $\alpha = 0.10$ andmake a decision based on the p-value approach. Explain fully.

22. A literary historian examines a newly discovered document possibly written by Oberon Theseus. The mean average sentence length of the surviving undisputed works of Oberon Theseus is 48.72 words. The historian counts words in sentences between five successive 101 periods in the document in question to obtain a mean average sentence length of 39.46 words with standard deviation 7.45 words. (Thus the sample size is five.)

a. Determine if these data provide sufficient evidence, at the 1% level of significance, to conclude that the mean average sentence length in the document is less than 48.72.
b. Estimate the p-value of the test.
c. Based on the answers to parts (a) and (b), state whether or not it is likely that the document was written by Oberon Theseus.

Answers

1. a. $Z \leq -1.645$
   b. $T \leq -2.571 \text{ or } T \geq 2.571$
   c. $T \geq 1.319$
   d. $Z \leq -1645 \text{ or } Z \geq 1.645$
2.
3. a. $T \leq -0.855$
   b. $Z \leq -1.645$
   c. $T \leq -2.201 \text{ or } T \geq 2.201$
   d. $T \geq 3.435$
4.
5. a. $T = -2.690,\ df = 19,\ -t_{0.005} = -2.861,$ do not reject H_0
   b. $0.01 < p - value < 0.02,\ \alpha = 0.01,$ do not reject H_0
6.
7. a. $T = 2.398,\ df = 7,\ t_{0.05} = 1.895,$ reject H_0
   b. $0.01 < p - value < 0.025,\ \alpha = 0.05,$ reject H_0
8.
9. $T = -7.560,\ df = 12,\ -t_{0.10} = -1.356,$ reject H_0
10.
11. $T = -7.076,\ df = 5,\ -t_{0.0005} = -6.869,$ reject H_0
12.
13. a. $T = -1.483,\ df = 14,\ -t_{0.05} = -1.761,$ do not reject H_0
    b. $T = -1.483,\ df = 14,\ -t_{0.10} = -1.345,$ reject H_0
14.
15. a. $T = 2.069,\ df = 6,\ t_{0.10} = 1.44,$ reject H_0
    b. $T = 2.069,\ df = 6,\ t_{0.05} = 1.943,$ reject H_0
16.
17. $T = 4.472,\ df = 4,\ t_{0.10} = 1.533,$ reject H_0
18.
19. $T = 0.798,\ df = 24,\ t_{0.10} = 1.318,$ do not reject H_0
20.
21. a. $T = -1.773,\ df = 4,\ -t_{0.05} = -2.132,$ do not reject H_0
    b. $0.05 < p - value < 0.10$
    c. $\alpha = 0.05,$ do not reject H_0

8.5: Large Sample Tests for a Population Proportion

Basic

On all exercises for this section you may assume that the sample is sufficiently large for the relevant test to be validly performed.

1. Compute the value of the test statistic for each test using the information given.
   a. Testing $H_0 : p = 0.50 \text{ vs } H_a : p > 0.50,\ n = 360,\ \hat{p} = 0.56.$
   b. Testing $H_0 : p = 0.50 \text{ vs } H_a : p \neq 0.50,\ n = 360,\ \hat{p} = 0.56.$
   c. Testing $H_0 : p = 0.37 \text{ vs } H_a : p < 0.37,\ n = 1200,\ \hat{p} = 0.35.$

2. Compute the value of the test statistic for each test using the information given.
   a. Testing $H_0 : p = 0.72$ *vs* $H_a : p < 0.72,\ n = 2100,\ \hat{p} = 0.71$.
   b. Testing $H_0 : p = 0.83$ *vs* $H_a : p \neq 0.83,\ n = 500,\ \hat{p} = 0.86$.
   c. Testing $H_0 : p = 0.22$ *vs* $H_a : p < 0.22,\ n = 750,\ \hat{p} = 0.18$.
3. For each part of Exercise 1 construct the rejection region for the test for $\alpha = 0.05$ and make the decision based on your answer to that part of the exercise.
4. For each part of Exercise 2 construct the rejection region for the test for $\alpha = 0.05$ and make the decision based on your answer to that part of the exercise.
5. For each part of Exercise 1 compute the observed significance (p-value) of the test and compare it to $\alpha = 0.05$ in order to make the decision by the p-value approach to hypothesis testing.
6. For each part of Exercise 2 compute the observed significance (p-value) of the test and compare it to $\alpha = 0.05$ in order to make the decision by the p-value approach to hypothesis testing.
7. Perform the indicated test of hypotheses using the critical value approach.
   a. Testing $H_0 : p = 0.55$ *vs* $H_a : p > 0.55$ @ $\alpha = 0.05,\ n = 300,\ \hat{p} = 0.60$.
   b. Testing $H_0 : p = 0.47$ *vs* $H_a : p \neq 0.47$ @ $\alpha = 0.01,\ n = 9750,\ \hat{p} = 0.46$.
8. Perform the indicated test of hypotheses using the critical value approach.
   a. Testing $H_0 : p = 0.15$ *vs* $H_a : p \neq 0.15$ @ $\alpha = 0.001,\ n = 1600,\ \hat{p} = 0.18$.
   b. Testing $H_0 : p = 0.90$ *vs* $H_a : p > 0.90$ @ $\alpha = 0.01,\ n = 1100,\ \hat{p} = 0.91$.
9. Perform the indicated test of hypotheses using the p-value approach.
   a. Testing $H_0 : p = 0.37$ *vs* $H_a : p \neq 0.37$ @ $\alpha = 0.005,\ n = 1300,\ \hat{p} = 0.40$.
   b. Testing $H_0 : p = 0.94$ *vs* $H_a : p > 0.94$ @ $\alpha = 0.05,\ n = 1200,\ \hat{p} = 0.96$.
10. Perform the indicated test of hypotheses using the p-value approach.
    a. Testing $H_0 : p = 0.25$ *vs* $H_a : p < 0.25$ @ $\alpha = 0.10,\ n = 850,\ \hat{p} = 0.23$.
    b. Testing $H_0 : p = 0.33$ *vs* $H_a : p \neq 0.33$ @ $\alpha = 0.05,\ n = 1100,\ \hat{p} = 0.30$.

Applications

11. Five years ago 3.9% of children in a certain region lived with someone other than a parent. A sociologist wishes to test whether the current proportion is different. Perform the relevant test at the 5% level of significance using the following data: in a random sample of $2,759$ children, 119 lived with someone other than a parent.
12. The government of a particular country reports its literacy rate as 52%. A nongovernmental organization believes it to be less. The organization takes a random sample of 600 inhabitants and obtains a literacy rate of 42%. Perform the relevant test at the 0.5% (one-half of 1%) level of significance.
13. Two years ago 72% of household in a certain county regularly participated in recycling household waste. The county government wishes to investigate whether that proportion has increased after an intensive campaign promoting recycling. In a survey of 900 households, 674 regularly participate in recycling. Perform the relevant test at the 10% level of significance.
14. Prior to a special advertising campaign, 23% of all adults recognized a particular company's logo. At the close of the campaign the marketing department commissioned a survey in which 311 of $1,200$ randomly selected adults recognized the logo. Determine, at the 1% level of significance, whether the data provide sufficient evidence to conclude that more than 23% of all adults now recognize the company's logo.
15. A report five years ago stated that 35.5% of all state-owned bridges in a particular state were "deficient." An advocacy group took a random sample of 100 state-owned bridges in the state and found 33 to be currently rated as being "deficient." Test whether the current proportion of bridges in such condition is 35.5% versus the alternative that it is different from 35.5%, at the 10% level of significance.
16. In the previous year the proportion of deposits in checking accounts at a certain bank that were made electronically was 45%. The bank wishes to determine if the proportion is higher this year. It examined $20,000$ deposit records and found that $9,217$ were electronic. Determine, at the 1% level of significance, whether the data provide sufficient evidence to conclude that more than 45% of all deposits to checking accounts are now being made electronically.

17. According to the Federal Poverty Measure 12% of the U.S. population lives in poverty. The governor of a certain state believes that the proportion there is lower. In a sample of size $1,550,163$ were impoverished according to the federal measure.
    a. Test whether the true proportion of the state's population that is impoverished is less than 12%, at the 5% level of significance.
    b. Compute the observed significance of the test.
18. An insurance company states that it settles 85% of all life insurance claims within 30 days. A consumer group asks the state insurance commission to investigate. In a sample of 250 life insurance claims, 203 were settled within 30 days.
    a. Test whether the true proportion of all life insurance claims made to this company that are settled within 30 days is less than 85%, at the 5% level of significance.
    b. Compute the observed significance of the test.
19. A special interest group asserts that 90% of all smokers began smoking before age 18. In a sample of 850 smokers, 687 began smoking before age 18.
    a. Test whether the true proportion of all smokers who began smoking before age 18 is less than 90%, at the 1% level of significance.
    b. Compute the observed significance of the test.
20. In the past, 68% of a garage's business was with former patrons. The owner of the garage samples 200 repair invoices and finds that for only 114 of them the patron was a repeat customer.
    a. Test whether the true proportion of all current business that is with repeat customers is less than 68%, at the 1% level of significance.
    b. Compute the observed significance of the test.

Additional Exercises

21. A rule of thumb is that for working individuals one-quarter of household income should be spent on housing. A financial advisor believes that the average proportion of income spent on housing is more than 0.25. In a sample of 30 households, the mean proportion of household income spent on housing was 0.285 with a standard deviation of 0.063. Perform the relevant test of hypotheses at the 1% level of significance. Hint: This exercise could have been presented in an earlier section.
22. Ice cream is legally required to contain at least 10% milk fat by weight. The manufacturer of an economy ice cream wishes to be close to the legal limit, hence produces its ice cream with a target proportion of 0.106 milk fat. A sample of five containers yielded a mean proportion of 0.094 milk fat with standard deviation 0.002. Test the null hypothesis that the mean proportion of milk fat in all containers is 0.106 against the alternative that it is less than 0.106, at the 10% level of significance. Assume that the proportion of milk fat in containers is normally distributed. Hint: This exercise could have been presented in an earlier section.

Large Data Set Exercises

Large Data Sets missing

23. Large **Data Sets 4 and 4A** list the results of 500 tosses of a die. Let p denote the proportion of all tosses of this die that would result in a five. Use the sample data to test the hypothesis that p is different from $1/6$, at the 20% level of significance.
24. Large **Data Set 6** records results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate. Use the full data set (400 observations) to test the hypothesis that the proportion p of all voters who prefer Candidate A exceeds 0.35. Test at the 10% level of significance.
25. Lines 2 through 536 in Large **Data Set 11** is a sample of 535 real estate sales in a certain region in 2008. Those that were foreclosure sales are identified with a 1 in the second column. Use these data to test, at the 10% level of significance, the hypothesis that the proportion p of all real estate sales in this region in 2008 that were foreclosure sales was less than 25%. (The null hypothesis is $H_0 : p = 0.25$).
26. Lines 537 through 1106 in Large **Data Set 11** is a sample of 570 real estate sales in a certain region in 2010. Those that were foreclosure sales are identified with a 1 in the second column. Use these data to test, at the 5% level of significance, the hypothesis that the proportion p of all real estate sales in this region in 2010 that were foreclosure sales was greater than 23%. (The null hypothesis is $H_0 : p = 0.25$).

Answers

1. a. $Z = 2.277$
   b. $Z = 2.277$
   c. $Z = -1.435$
2.
3. a. $Z \geq 1.645$; reject H_0
   b. $Z \leq -1.96$ *or* $Z \geq 1.96$; reject H_0
   c. $Z \leq -1.645$; do not reject H_0
4.
5. a. $p - value = 0.0116,\ \alpha = 0.05$; reject H_0
   b. $p - value = 0.0232,\ \alpha = 0.05$; reject H_0
   c. $p - value = 0.0749,\ \alpha = 0.05$; do not reject H_0
6.
7. a. $Z = 1.74,\ z_{0.05} = 1.645$; reject H_0
   b. $Z = -1.98,\ -z_{0.005} = -2.576$; do not reject H_0
8.
9. a. $Z = 2.24,\ p - value = 0.025, \alpha = 0.005$; do not reject H_0
   b. $Z = 2.92,\ p - value = 0.0018, \alpha = 0.05$; reject H_0
10.
11. $Z = 1.11,\ z_{0.025} = 1.96$; do not reject H_0
12.
13. $Z = 1.93,\ z_{0.10} = 1.28$; reject H_0
14.
15. $Z = -0.523,\ \pm z_{0.05} = \pm 1.645$; do not reject H_0
16.
17. a. $Z = -1.798,\ -z_{0.05} = -1.645$; reject H_0
    b. $p - value = 0.0359$
18.
19. a. $Z = -8.92,\ -z_{0.01} = -2.33$; reject H_0
    b. $p - value \approx 0$
20.
21. $Z = 3.04,\ z_{0.01} = 2.33$; reject H_0
22.
23. $H_0 : p = 1/6$ *vs* $H_a : p \neq 1/6$. Test Statistic: $Z = -0.76$. Rejection Region: $(-\infty, -1.28] \cup [1.28, \infty)$. Decision: Fail to reject H_0.
24.
25. $H_0 : p = 0.25$ *vs* $H_a : p < 0.25$. Test Statistic: $Z = -1.17$. Rejection Region: $(-\infty, -1.28]$. Decision: Fail to reject H_0.

Contributor

- Anonymous

This page titled 8.E: Testing Hypotheses (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **8.E: Testing Hypotheses (Exercises)** has no license indicated.

CHAPTER OVERVIEW

9: Two-Sample Problems

The previous two chapters treated the questions of estimating and making inferences about a parameter of a single population. In this chapter we consider a comparison of parameters that belong to two different populations. For example, we might wish to compare the average income of all adults in one region of the country with the average income of those in another region, or we might wish to compare the proportion of all men who are vegetarians with the proportion of all women who are vegetarians. We will study construction of confidence intervals and tests of hypotheses in four situations, depending on the parameter of interest, the sizes of the samples drawn from each of the populations, and the method of sampling. We also examine sample size considerations.

9.1: Two Population Proportions
9.2: Two Population Means - Independent Samples
9.2.1: Large, Independent Samples
9.2.2: Small, Independent Samples
9.3: Two Population Means - Paired Samples
9.4: Sample Size Considerations
9.E: Two-Sample Problems (Exercises)

This page titled 9: Two-Sample Problems is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

9.1: Two Population Proportions

Learning Objectives

- To learn how to construct a confidence interval for the difference in the proportions of two distinct populations that have a particular characteristic of interest.
- To learn how to perform a test of hypotheses concerning the difference in the proportions of two distinct populations that have a particular characteristic of interest.

Suppose we wish to compare the proportions of two populations that have a specific characteristic, such as the proportion of men who are left-handed compared to the proportion of women who are left-handed. Figure **9.1.1** illustrates the conceptual framework of our investigation. Each population is divided into two groups, the group of elements that have the characteristic of interest (for example, being left-handed) and the group of elements that do not. We arbitrarily label one population as Population **1** and the other as Population **2**, and subscript the proportion of each population that possesses the characteristic with the number **1** or **2** to tell them apart. We draw a random sample from Population **1** and label the sample statistic it yields with the subscript **1**. Without reference to the first sample we draw a sample from Population **2** and label its sample statistic with the subscript **2**.

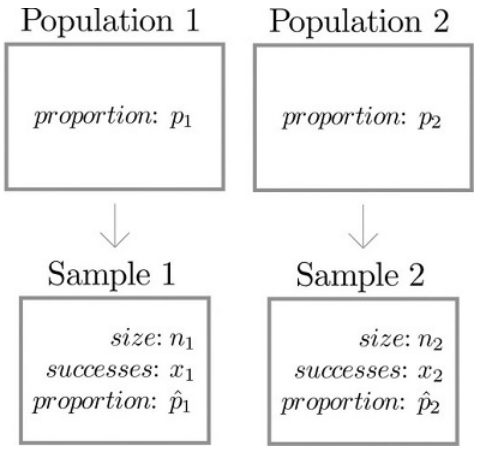


Figure **9.1.1**: Independent Sampling from Two Populations In Order to Compare Proportions

Our goal is to use the information in the samples to estimate the difference $p_1 - p_2$ in the two population proportions and to make statistically valid inferences about it.

Confidence Intervals

Since the sample proportion $\hat{p}_1$ computed using the sample drawn from Population **1** is a good estimator of population proportion p_1 of Population **1** and the sample proportion $\hat{p}_2$ computed using the sample drawn from Population **2** is a good estimator of population proportion p_2 of Population **2**, a reasonable point estimate of the difference $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. In order to widen this point estimate into a confidence interval we suppose that both samples are large, as described in Section 7.3 and repeated below. If so, then the following formula for a confidence interval for $p_1 - p_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The samples must be independent, and *each* sample must be large: each of the intervals

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right]$$

and

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

must lie wholly within the interval $[0, 1]$.

Example 9.1.1

The department of code enforcement of a county government issues permits to general contractors to work on residential projects. For each permit issued, the department inspects the result of the project and gives a "pass" or "fail" rating. A failed project must be re-inspected until it receives a pass rating. The department had been frustrated by the high cost of re-inspection and decided to publish the inspection records of all contractors on the web. It was hoped that public access to the records would lower the re-inspection rate. A year after the web access was made public, two samples of records were randomly selected. One sample was selected from the pool of records before the web publication and one after. The proportion of projects that passed on the first inspection was noted for each sample. The results are summarized below. Construct a point estimate and a 90% confidence interval for the difference in the passing rate on first inspection between the two time periods.

No public web access	$n_1 = 500 \;\; \hat{p}_1 = 0.67$
Public web access	$n_2 = 100 \;\; \hat{p}_2 = 0.80$

Solution

The point estimate of $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 = 0.67 - 0.80 = -0.13$$

Because the "No public web access" population was labeled as Population 1 and the "Public web access" population was labeled as Population 2, in words this means that we estimate that the proportion of projects that passed on the first inspection increased by 13 percentage points after records were posted on the web.

The sample sizes are sufficiently large for constructing a confidence interval since for sample 1:

$$3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 3\sqrt{\frac{(0.67)(0.33)}{500}} = 0.06$$

so that

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right] = [0.67 - 0.06, 0.67 + 0.06] = [0.61, 0.73] \subset [0, 1]$$

and for sample 2:

$$3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 3\sqrt{\frac{(0.8)(0.2)}{100}} = 0.12$$

so that

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right] = [0.8 - 0.12, 0.8 + 0.12] = [0.68, 0.92] \subset [0, 1]$$

To apply the formula for the confidence interval, we first observe that the 90% confidence level means that $\alpha = 1 - 0.90 = 0.10$ so that $z_{\alpha/2} = z_{0.05}$. From Figure 7.1.6 we read directly that $z_{0.05} = 1.645$. Thus the desired confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{9.1.1}$$

$$= 0.13 \pm 1.645\sqrt{\frac{(0.67)(0.33)}{500} + \frac{(0.8)(0.2)}{100}} \tag{9.1.2}$$

$$= -0.13 \pm 0.07 \tag{9.1.3}$$

The 90% confidence interval is $[-0.20, -0.06]$. We are 90% confident that the difference in the population proportions lies in the interval $[-0.20, -0.06]$, in the sense that in repeated sampling 90% of all intervals constructed from the sample data in this manner will contain $p_1 - p_2$. Taking into account the labeling of the two populations, this means that we are 90% confident that the proportion of projects that pass on the first inspection is between 6 and 20 percentage points higher after public access to the records than before.

Hypothesis Testing

In hypothesis tests concerning the relative sizes of the proportions p_1 and p_2 of two populations that possess a particular characteristic, the null and alternative hypotheses will always be expressed in terms of the difference of the two population proportions. Hence the null hypothesis is always written

$$H_0 : p_1 - p_2 = D_0$$

The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : p_1 - p_2 < D_0$	Left-tailed
$H_a : p_1 - p_2 > D_0$	Right-tailed
$H_a : p_1 - p_2 \neq D_0$	Two-tailed

As long as the samples are independent and both are large the following formula for the standardized test statistic is valid, and it has the standard normal distribution.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Proportions

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

The test statistic has the standard normal distribution.

The samples must be independent, and each sample must be large: each of the intervals

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right]$$

and

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

must lie wholly within the interval $[0, 1]$.

Example 9.1.2

Using the data of Example 9.1.1, test whether there is sufficient evidence to conclude that public web access to the inspection records has increased the proportion of projects that passed on the first inspection by more than 5 percentage points. Use the critical value approach at the 10% level of significance.

Solution

- **Step 1**. Taking into account the labeling of the populations an increase in passing rate at the first inspection by more than 5 percentage points after public access on the web may be expressed as $p_2 > p_1 + 0.05$, which by algebra is the same as $p_1 - p_2 < -0.05$. This is the alternative hypothesis. Since the null hypothesis is always expressed as an equality, with the same

number on the right as is in the alternative hypothesis, the test is

$$H_0 : p_1 - p_2 = -0.05$$
vs.
$$H_a : p_1 - p_2 < -0.05 \ @ \ \alpha = 0.10$$

- **Step 2**. Since the test is with respect to a difference in population proportions the test statistic is

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - D_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

- **Step 3**. Inserting the values given in Example **9.1.1** and the value $D_0 = -0.05$into the formula for the test statistic gives

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - D_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} = \frac{(-0.13) - (-0.05)}{\sqrt{\frac{(0.67)(0.33)}{500} + \frac{(0.8)(0.2)}{100}}} = -1.770$$

- **Step 4**. Since the symbol in H_a is "<" this is a left-tailed test, so there is a single critical value, $z_\alpha = -z_{0.10}$. From the last row in Figure 7.1.6 $z_{0.10} = 1.282$, so $-z_{0.10} = -1.282$. The rejection region is $(-\infty, -1.282]$.
- **Step 5**. As shown in Figure **9.1.2** the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the **10%** level of significance, to conclude that the rate of passing on the first inspection has increased by more than **5** percentage points since records were publicly posted on the web.

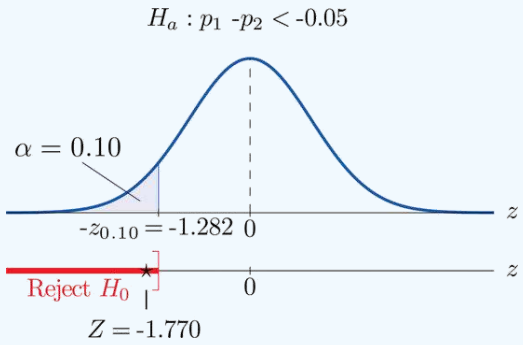


Figure **9.1.2**: Rejection Region and Test Statistic for "Example **9.1.2**"

Example 9.1.3

Perform the test of Example **9.1.2** using the p-value approach.

Solution

The first three steps are identical to those in Example **9.1.2**

- **Step 4**. Because the test is left-tailed the observed significance or p-value of the test is just the area of the left tail of the standard normal distribution that is cut off by the test statistic $Z = -1.770$. From Figure 7.1.5 the area of the left tail determined by -1.77 is **0.0384**. The p-value is **0.0384**.
- **Step 5**. Since the p-value **0.0384** is less than $\alpha = 0.10$, the decision is to reject the null hypothesis: The data provide sufficient evidence, at the **10%** level of significance, to conclude that the rate of passing on the first inspection has increased by more than **5** percentage points since records were publicly posted on the web.

Finally a common misuse of the formulas given in this section must be mentioned. Suppose a large pre-election survey of potential voters is conducted. Each person surveyed is asked to express a preference between, say, Candidate A and Candidate B. (Perhaps "no preference" or "other" are also choices, but that is not important.) In such a survey, estimators $\hat{p}_A$ and $\hat{p}B$ of p_A and p_B can be calculated. It is important to realize, however, that these two estimators were not calculated from two independent samples. While $\hat{p}A - \hat{p}_B$ may be a reasonable estimator of $p_A - p_B$, the formulas for confidence intervals and for the standardized test statistic given in this section are not valid for data obtained in this manner.

Key Takeaway

- A confidence interval for the difference in two population proportions is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population proportion is used to test hypotheses concerning the difference between two population proportions. The only difference is in the formula for the standardized test statistic.

This page titled 9.1: Two Population Proportions is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **9.4: Comparison of Two Population Proportions** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.2: Two Population Means - Independent Samples

9.2: Two Population Means - Independent Samples is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.2.1: Large, Independent Samples

Learning Objectives

- To understand the logical framework for estimating the difference between the means of two distinct populations and performing tests of hypotheses concerning those means.
- To learn how to construct a confidence interval for the difference in the means of two distinct populations using large, independent samples.
- To learn how to perform a test of hypotheses concerning the difference between the means of two distinct populations using large, independent samples.

Suppose we wish to compare the means of two distinct populations. Figure **9.2.1.1** illustrates the conceptual framework of our investigation in this and the next section. Each population has a mean and a standard deviation. We arbitrarily label one population as Population **1** and the other as Population **2**, and subscript the parameters with the numbers **1** and **2** to tell them apart. We draw a random sample from Population **1** and label the sample statistics it yields with the subscript **1**. Without reference to the first sample we draw a sample from Population **2** and label its sample statistics with the subscript **2**.

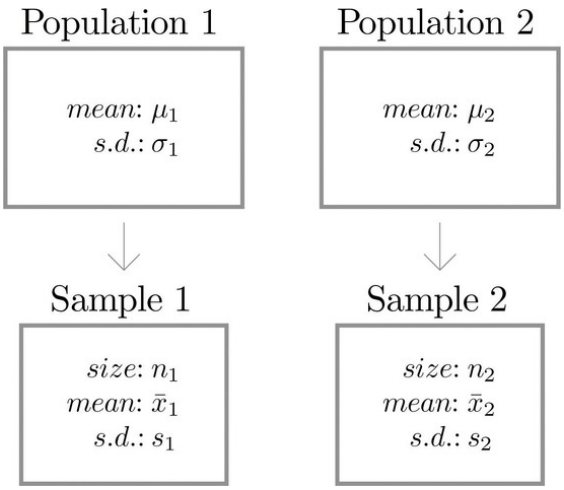


Figure **9.2.1.1**: Independent Sampling from Two Populations

Definition: Independence

Samples from two distinct populations are *independent* if each one is drawn without reference to the other, and has no connection with the other.

Our goal is to use the information in the samples to estimate the difference $\mu_1 - \mu_2$ in the means of the two populations and to make statistically valid inferences about it.

Confidence Intervals

Since the mean $x - 1$ of the sample drawn from Population **1** is a good estimator of μ_1 and the mean $x - 2$ of the sample drawn from Population **2** is a good estimator of μ_2, a reasonable point estimate of the difference $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$. In order to widen this point estimate into a confidence interval, we first suppose that both samples are large, that is, that both $n_1 \geq 30$ and $n_2 \geq 30$. If so, then the following formula for a confidence interval for $\mu_1 - \mu_2$ is valid. The symbols s_1^2 and s_2^2 denote the squares of s_1 and s_2. (In the relatively rare case that both population standard deviations σ_1 and σ_2 are known they would be used instead of the sample standard deviations.)

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Large, Independent Samples

The samples must be independent, and *each* sample must be large:

Example 9.2.1.1

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company 1	Company 2
$n_1 = 174$	$n_2 = 355$
$x-1 = 3.51$	$x-2 = 3.24$
$s_1 = 0.51$	$s_2 = 0.52$

Construct a point estimate and a 99% confidence interval for $\mu_1 - \mu_2$, the difference in average satisfaction levels of customers of the two companies as measured on this five-point scale.

Solution

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x_1} - \bar{x_2} = 3.51 - 3.24 = 0.27$$

In words, we estimate that the average customer satisfaction level for Company 1 is 0.27 points higher on this five-point scale than it is for Company 2.

To apply the formula for the confidence interval, proceed exactly as was done in Chapter 7. The 99% confidence level means that $\alpha = 1 - 0.99 = 0.01$ so that $z_{\alpha/2} = z_{0.005}$. From Figure 7.1.6 "Critical Values of " we read directly that $z_{0.005} = 2.576$. Thus

$$(\bar{x_1} - \bar{x_2}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.27 \pm 2.576\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}} = 0.27 \pm 0.12$$

We are 99% confident that the difference in the population means lies in the interval $[0.15, 0.39]$, in the sense that in repeated sampling 99% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. In the context of the problem we say we are 99% confident that the average level of customer satisfaction for Company 1 is between 0.15 and 0.39 points higher, on this five-point scale, than that for Company 2.

Hypothesis Testing

Hypotheses concerning the relative sizes of the means of two populations are tested using the same critical value and p-value procedures that were used in the case of a single population. All that is needed is to know how to express the null and alternative hypotheses and to know the formula for the standardized test statistic and the distribution that it follows.

The null and alternative hypotheses will always be expressed in terms of the difference of the two population means. Thus the null hypothesis will always be written

$$H_0 : \mu_1 - \mu_2 = D_0$$

where D_0 is a number that is deduced from the statement of the situation. As was the case with a single population the alternative hypothesis can take one of the three forms, with the same terminology:

Form of H*a*	Terminology
$H_a : \mu_1 - \mu_2 < D_0$	Left-tailed
$H_a : \mu_1 - \mu_2 > D_0$	Right-tailed
$H_a : \mu_1 - \mu_2 \neq D_0$	Two-tailed

As long as the samples are independent and both are large the following formula for the standardized test statistic is valid, and it has the standard normal distribution. (In the relatively rare case that both population standard deviations σ_1 and σ_2 are known they would be used instead of the sample standard deviations.)

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Large, Independent Samples

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The test statistic has the standard normal distribution.

The samples must be independent, and each sample must be large: $n_1 \geq 30$ and $n_2 \geq 30$.

Example 9.2.1.2

Refer to Example **9.2.1.1** concerning the mean satisfaction levels of customers of two competing cable television companies. Test at the **1%** level of significance whether the data provide sufficient evidence to conclude that Company **1** has a higher mean satisfaction rating than does Company **2**. Use the critical value approach.

Solution:

- **Step 1**. If the mean satisfaction levels μ_1 and μ_2 are the same then $\mu_1 = \mu_2$, but we always express the null hypothesis in terms of the difference between μ_1 and μ_2, hence H_0 is $\mu_1 - \mu_2 = 0$. To say that the mean customer satisfaction for Company **1** is higher than that for Company **2** means that $\mu_1 > \mu_2$, which in terms of their difference is $\mu_1 - \mu_2 > 0$. The test is therefore

$$H_0 : \mu_1 - \mu_2 = 0$$

$$vs.$$

$$H_a : \mu_1 - \mu_2 > 0 \ @ \ \alpha = 0.01$$

- **Step 2**. Since the samples are independent and both are large the test statistic is

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.51 - 3.24) - 0}{\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}}} = 5.684$$

- **Step 4**. Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $z_\alpha = z_{0.01}$, which from the last line in Figure 7.1.6 "Critical Values of " we read off as **2.326**. The rejection region is $[2.326, \infty)$.

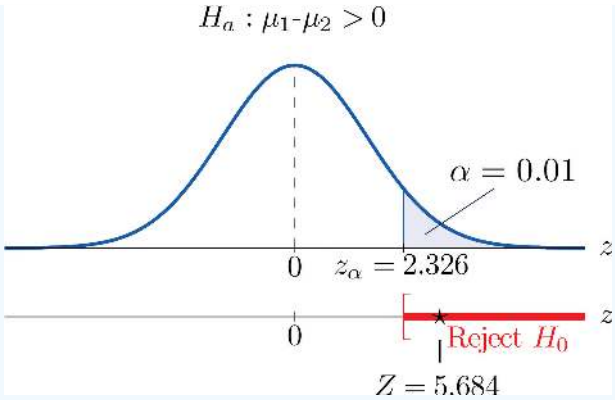


***Figure* 9.2.1.2**: *Rejection Region and Test Statistic for Example* **9.2.1.2**

- **Step 5**. As shown in Figure **9.2.1.2** the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the **1%** level of significance, to conclude that the mean customer satisfaction for Company **1** is higher than that for Company **2**.

Example 9.2.1.3

Perform the test of Example **9.2.1.2** using the p-value approach.

Solution:

The first three steps are identical to those in Example **9.2.1.2**

- **Step 4**. The observed significance or p-value of the test is the area of the right tail of the standard normal distribution that is cut off by the test statistic $Z = 5.684$. The number **5.684** is too large to appear in Figure 7.1.5, which means that the area of the left tail that it cuts off is **1.0000** to four decimal places. The area that we seek, the area of the right tail, is therefore $1 - 1.0000 = 0.0000$ to four decimal places. See Figure **9.2.1.3**. That is, p-value=**0.0000** to four decimal places. (The actual value is approximately **0.000000007**.)

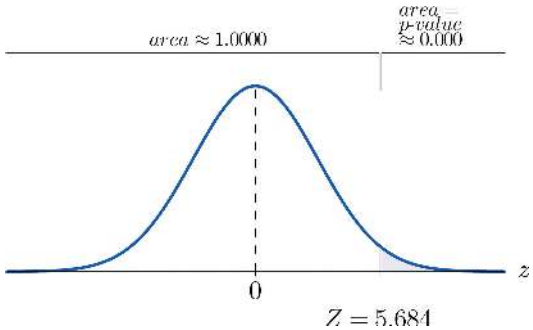


Figure **9.2.1.3**: ***P*-*Value for Example* 9.2.1.3**

- **Step 5**. Since $0.0000 < 0.01$, $p - value < \alpha$ so the decision is to reject the null hypothesis:

  The data provide sufficient evidence, at the **1%** level of significance, to conclude that the mean customer satisfaction for Company **1** is higher than that for Company **2**.

Key Takeaway

- A point estimate for the difference in two population means is simply the difference in the corresponding sample means.
- In the context of estimating or testing hypotheses concerning two population means, “large” samples means that both samples are large.

- A confidence interval for the difference in two population means is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population mean is used to test hypotheses concerning the difference between two population means. The only difference is in the formula for the standardized test statistic.

This page titled 9.2.1: Large, Independent Samples is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **9.1: Comparison of Two Population Means- Large, Independent Samples** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.2.2: Small, Independent Samples

Learning Objectives

- To learn how to construct a confidence interval for the difference in the means of two distinct populations using small, independent samples.
- To learn how to perform a test of hypotheses concerning the difference between the means of two distinct populations using small, independent samples.

When one or the other of the sample sizes is small, as is often the case in practice, the Central Limit Theorem does not apply. We must then impose conditions on the population to give statistical validity to the test procedure. We will assume that both populations from which the samples are taken have a normal probability distribution and that their standard deviations are equal.

Confidence Intervals

When the two populations are normally distributed and have equal standard deviations, the following formula for a confidence interval for $\mu_1 - \mu_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Small, Independent Samples

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{9.2.2.1}$$

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

The number of degrees of freedom is

$$df = n_1 + n_2 - 2.$$

The samples must be independent, the populations must be normal, and the population standard deviations must be equal. "Small" samples means that either $n_1 < 30$ or $n_2 < 30$.

The quantity s_p^2 is called the **pooled sample variance**. It is a weighted average of the two estimates s_1^2 and s_2^2 of the common variance $\sigma_1^2 = \sigma_2^2$ of the two populations.

Example 9.2.2.1

A software company markets a new computer game with two experimental packaging designs. Design 1 is sent to 11 stores; their average sales the first month is 52 units with sample standard deviation 12 units. Design 2 is sent to 6 stores; their average sales the first month is 46 units with sample standard deviation 10 units. Construct a point estimate and a 95% confidence interval for the difference in average monthly sales between the two package designs.

Solution

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 = 52 - 46 = 6$$

In words, we estimate that the average monthly sales for Design 1 is 6 units more per month than the average monthly sales for Design 2.

To apply the formula for the confidence interval (Equation 9.2.2.1), we must find $t_{\alpha/2}$. The 95% confidence level means that $\alpha = 1 - 0.95 = 0.05$ so that $t_{\alpha/2} = t_{0.025}$. From Figure 7.1.6, in the row with the heading $df = 11 + 6 - 2 = 15$ we read that $t_{0.025} = 2.131$. From the formula for the pooled sample variance we compute

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(10)(12)^2 + (5)(10)^2}{15} = 129.\bar{3}$$

Thus

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha}/2\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 6 \pm (2.131)\sqrt{129.\bar{3}\left(\frac{1}{11}+\frac{1}{6}\right)} \approx 6 \pm 12.3$$

We are 95% confident that the difference in the population means lies in the interval $[-6.3, 18.3]$, in the sense that in repeated sampling 95% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. Because the interval contains both positive and negative values the statement in the context of the problem is that we are 95% confident that the average monthly sales for Design 1 is between 18.3 units higher and 6.3 units lower than the average monthly sales for Design 2.

Hypothesis Testing

Testing hypotheses concerning the difference of two population means using small samples is done precisely as it is done for large samples, using the following standardized test statistic. The same conditions on the populations that were required for constructing a confidence interval for the difference of the means must also be met when hypotheses are tested.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Small, Independent Samples

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

The test statistic has Student's t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom.

The samples must be independent, the populations must be normal, and the population standard deviations must be equal. "Small" samples means that either $n_1 < 30$ or $n_2 < 30$.

Example 9.2.2.2

Refer to Example 9.2.2.1 concerning the mean sales per month for the same computer game but sold with two package designs. Test at the 1% level of significance whether the data provide sufficient evidence to conclude that the mean sales per month of the two designs are different. Use the critical value approach.

Solution

- **Step 1**. The relevant test is

$$H_0 : \mu_1 - \mu_2 = 0$$

vs.

$$H_a : \mu_1 - \mu_2 \neq 0 \ @ \ \alpha = 0.01$$

- **Step 2**. Since the samples are independent and at least one is less than 30 the test statistic is

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

which has Student's t-distribution with $df = 11 + 6 - 2 = 15$ degrees of freedom.

- **Step 3**. Inserting the data and the value $D_0 = 0$ into the formula for the test statistic gives

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(52 - 46) - 0}{\sqrt{129.\bar{3}\left(\frac{1}{11} + \frac{1}{6}\right)}} = 1.040$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm t_{\alpha/2} = \pm t_{0.005}$. From the row in Figure 7.1.6 with the heading $df = 15$ we read off $t_{0.005} = 2.947$. The rejection region is $(-\infty, -2.947] \cup [2.947, \infty)$.

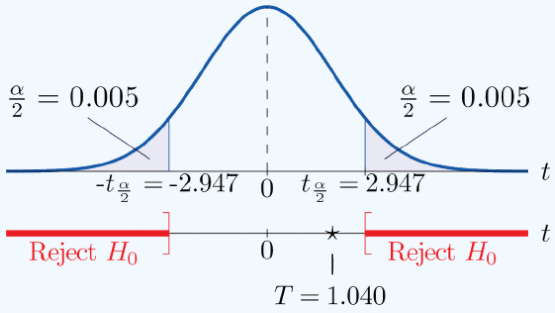


Figure 9.2.2.1: Rejection Region and Test Statistic for "Example 9.2.2.2"

- **Step 5**. As shown in Figure 9.2.2.1 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

  The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the mean sales per month of the two designs are different.

Example 9.2.2.3

Perform the test of Example 9.2.2.2 using the p-value approach.

Solution

The first three steps are identical to those in Example 9.2.2.2.

- **Step 4**. Because the test is two-tailed the observed significance or p-value of the test is the double of the area of the right tail of Student'st-distribution, with 15 degrees of freedom, that is cut off by the test statistic $T = 1.040$. We can only approximate this number. Looking in the row of Figure 7.1.6 headed $df = 15$, the number 1.040 is between the numbers 0.866 and 1.341, corresponding to $t_{0.200}$ and $t_{0.100}$. The area cut off by $t = 0.866$ is 0.200 and the area cut off by $t = 1.341$ is 0.100. Since 1.040 is between 0.866 and 1.341 the area it cuts off is between 0.200 and 0.100. Thus the p-value (since the area must be doubled) is between 0.400 and 0.200.
- **Step 5**. Since $p > 0.200 > 0.01$, $p > \alpha$, so the decision is not to reject the null hypothesis:

  The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the mean sales per month of the two designs are different.

Key Takeaway

- In the context of estimating or testing hypotheses concerning two population means, "small" samples means that at least one sample is small. In particular, even if one sample is of size 30 or more, if the other is of size less than 30 the formulas of this section must be used.
- A confidence interval for the difference in two population means is computed using a formula in the same fashion as was done for a single population mean.

This page titled 9.2.2: Small, Independent Samples is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **9.2: Comparison of Two Population Means - Small, Independent Samples** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.3: Two Population Means - Paired Samples

Learning Objectives

- To learn the distinction between independent samples and paired samples.
- To learn how to construct a confidence interval for the difference in the means of two distinct populations using paired samples.
- To learn how to perform a test of hypotheses concerning the difference in the means of two distinct populations using paired samples

Suppose chemical engineers wish to compare the fuel economy obtained by two different formulations of gasoline. Since fuel economy varies widely from car to car, if the mean fuel economy of two independent samples of vehicles run on the two types of fuel were compared, even if one formulation were better than the other the large variability from vehicle to vehicle might make any difference arising from difference in fuel difficult to detect. Just imagine one random sample having many more large vehicles than the other. Instead of independent random samples, it would make more sense to select pairs of cars of the same make and model and driven under similar circumstances, and compare the fuel economy of the two cars in each pair. Thus the data would look something like Table **9.3.1**, where the first car in each pair is operated on one formulation of the fuel (call it Type **1** gasoline) and the second car is operated on the second (call it Type **2** gasoline).

Table **9.3.1**: Fuel Economy of Pairs of Vehicles

Make and Model	Car 1	Car 2
Buick LaCrosse	17.0	17.0
Dodge Viper	13.2	12.9
Honda CR-Z	35.3	35.4
Hummer H 3	13.6	13.2
Lexus RX	32.7	32.5
Mazda CX-9	18.4	18.1
Saab 9-3	22.5	22.5
Toyota Corolla	26.8	26.7
Volvo XC 90	15.1	15.0

The first column of numbers form a sample from Population **1**, the population of all cars operated on Type **1** gasoline; the second column of numbers form a sample from Population **2**, the population of all cars operated on Type **2** gasoline. It would be incorrect to analyze the data using the formulas from the previous section, however, since the samples were not drawn independently. What is correct is to compute the difference in the numbers in each pair (subtracting in the same order each time) to obtain the third column of numbers as shown in Table **9.3.2** and treat the differences as the data. At this point, the new sample of differences $d_1 = 0.0, \cdots, d_9 = 0.1$ in the third column of Table **9.3.2** may be considered as a random sample of size $n = 9$ selected from a population with mean $\mu_d = \mu_1 - \mu_2$. This approach essentially transforms the paired two-sample problem into a one-sample problem as discussed in the previous two chapters.

Table **9.3.2**: Fuel Economy of Pairs of Vehicles

Make and Model	Car 1	Car 2	Difference
Buick LaCrosse	17.0	17.0	0.0
Dodge Viper	13.2	12.9	0.3
Honda CR-Z	35.3	35.4	−0.1
Hummer H 3	13.6	13.2	0.4
Lexus RX	32.7	32.5	0.2
Mazda CX-9	18.4	18.1	0.3
Saab 9-3	22.5	22.5	0.0
Toyota Corolla	26.8	26.7	0.1
Volvo XC 90	15.1	15.0	0.1

Note carefully that although it does not matter what order the subtraction is done, it must be done in the same order for all pairs. This is why there are both positive and negative quantities in the third column of numbers in Table **9.3.2**.

Confidence Intervals

When the population of differences is normally distributed the following formula for a confidence interval for $\mu_d = \mu_1 - \mu_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Paired Difference Samples

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

where there are n pairs, $\bar{d}$ is the mean and s_d is the standard deviation of their differences.

The number of degrees of freedom is

$$df = n - 1.$$

The population of differences must be *normally distributed*.

Example 9.3.1

Using the data in Table **9.3.1** construct a point estimate and a **95%** confidence interval for the difference in average fuel economy between cars operated on Type **1** gasoline and cars operated on Type **2** gasoline.

Solution

We have referred to the data in Table **9.3.1** because that is the way that the data are typically presented, but we emphasize that with paired sampling one immediately computes the differences, as given in Table **9.3.2**, and uses the differences as the data.

The mean and standard deviation of the differences are

$$\bar{d} = \frac{\sum d}{n} = \frac{1.3}{9} = 0.1\bar{4}$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{1}{n}(\sum d)^2}{n-1}} = \sqrt{\frac{0.41 - \frac{1}{9}(1.3)^2}{8}} = 0.1\bar{6}$$

The point estimate of $\mu_1 - \mu_2 = \mu_d$ is

$$\bar{d} = 0.14$$

In words, we estimate that the average fuel economy of cars using Type 1 gasoline is 0.14 mpg greater than the average fuel economy of cars using Type 2 gasoline.

To apply the formula for the confidence interval, we must find $t_{\alpha/2}$. The 95% confidence level means that $\alpha = 1 - 0.95 = 0.05$ so that $t_{\alpha/2} = t_{0.025}$. From Figure 7.1.6, in the row with the heading $df = 9 - 1 = 8$ we read that $t_{0.025} = 2.306$. Thus

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}} = 0.14 \pm 2.306\left(\frac{0.1\bar{6}}{\sqrt{9}}\right) \approx 0.14 \pm 0.13$$

We are 95% confident that the difference in the population means lies in the interval $[0.01, 0.27]$, in the sense that in repeated sampling 95% of all intervals constructed from the sample data in this manner will contain $\mu_d = \mu_1 - \mu_2$. Stated differently, we are 95% confident that mean fuel economy is between 0.01 and 0.27 mpg greater with Type 1 gasoline than with Type 2 gasoline.

Hypothesis Testing

Testing hypotheses concerning the difference of two population means using paired difference samples is done precisely as it is done for independent samples, although now the null and alternative hypotheses are expressed in terms of μ_d instead of $\mu_1 - \mu_2$. Thus the null hypothesis will always be written

$$H_0 : \mu_d = D_0$$

The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : \mu_d < D_0$	Left-tailed
$H_a : \mu_d > D_0$	Right-tailed
$H_a : \mu_d \neq D_0$	Two-tailed

The same conditions on the population of differences that was required for constructing a confidence interval for the difference of the means must also be met when hypotheses are tested. Here is the standardized test statistic that is used in the test.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Paired Difference Samples

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

where there are n pairs, $\bar{d}$ is the mean and s_d is the standard deviation of their differences.

The test statistic has Student's t-distribution with $df = n - 1$ degrees of freedom.

The population of differences must be normally distributed.

Example 9.3.2: using the critical value approach

Using the data of Table 9.3.2 test the hypothesis that mean fuel economy for Type 1 gasoline is greater than that for Type 2 gasoline against the null hypothesis that the two formulations of gasoline yield the same mean fuel economy. Test at the 5% level of significance using the critical value approach.

Solution

The only part of the table that we use is the third column, the differences.

- **Step 1**. Since the differences were computed in the order $\textbf{Type 1 mpg} - \textbf{Type 2 mpg}$, better fuel economy with Type 1 fuel corresponds to $\mu_d = \mu_1 - \mu_2 > 0$. Thus the test is

$$H_0 : \mu_d = 0$$
vs.
$$H_a : \mu_d > 0 \ @ \ \alpha = 0.05$$

(If the differences had been computed in the opposite order then the alternative hypotheses would have been $H_a : \mu_d < 0$.)

- **Step 2**. Since the sampling is in pairs the test statistic is

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

- **Step 3**. We have already computed $\bar{d}$ and s_d in the previous example. Inserting their values and $D_0 = 0$ into the formula for the test statistic gives

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} = \frac{0.1\bar{4}}{0.1\bar{6}/\sqrt{3}} = 2.600$$

- **Step 4**. Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $t_\alpha = t_{0.05}$ with 8 degrees of freedom, which from the row labeled $df = 8$ in Figure 7.1.6 we read off as 1.860. The rejection region is $[1.860, \infty)$.
- **Step 5**. As shown in Figure 9.3.1 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

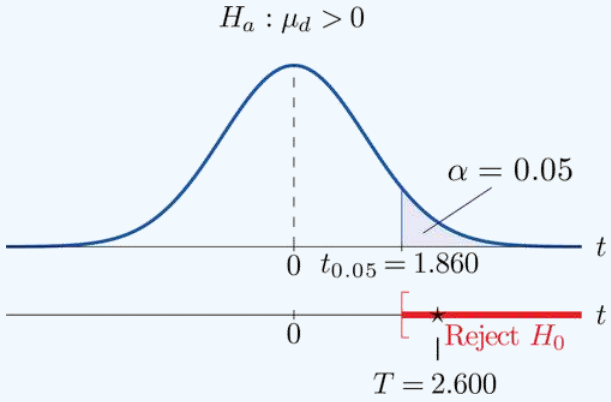


Figure 9.3.1: Rejection Region and Test Statistic for "Example 9.3.2"

The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean fuel economy provided by Type 1 gasoline is greater than that for Type 2 gasoline.

✔ Example 9.3.3: using the p-value approach

Perform the test in Example 9.3.2 using the p-value approach.

Solution

The first three steps are identical to those 9.3.2.

- **Step 4**. Because the test is one-tailed the observed significance or p-value of the test is just the area of the right tail of Student's t-distribution, with 8 degrees of freedom, that is cut off by the test statistic $T = 2.600$. We can only approximate this number. Looking in the row of Figure 7.1.6 headed $df = 8$, the number 2.600 is between the numbers 2.306 and 2.896, corresponding to $t_{0.025}$ and $t_{0.010}$. The area cut off by $t = 2.306$ is 0.025 and the area cut off by $t = 2.896$ is 0.010. Since 2.600 is between 2.306 and 2.896 the area it cuts off is between 0.025 and 0.010. Thus the p-value is between 0.025 and 0.010. In particular it is less than 0.025. See Figure 9.3.2.

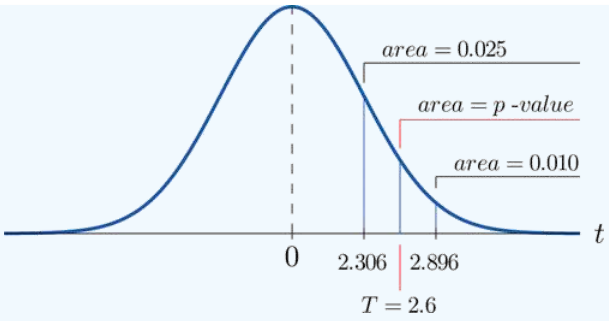


Figure 9.3.2: P-Value for "Example 9.3.3"

- **Step 5**. Since $0.025 < 0.05$, $p < \alpha$ so the decision is to reject the null hypothesis:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean fuel economy provided by Type 1 gasoline is greater than that for Type 2 gasoline.

The paired two-sample experiment is a very powerful study design. It bypasses many unwanted sources of "statistical noise" that might otherwise influence the outcome of the experiment, and focuses on the possible difference that might arise from the one factor of interest.

If the sample is large (meaning that $n \geq 30$) then in the formula for the confidence interval we may replace $t_{\alpha/2}$ by $z_{\alpha/2}$. For hypothesis testing when the number of pairs is at least 30, we may use the same statistic as for small samples for hypothesis testing, except now it follows a standard normal distribution, so we use the last line of Figure 7.1.6 to compute critical values, and p-values can be computed exactly with Figure 7.1.5, not merely estimated using Figure 7.1.6.

Key Takeaway

- When the data are collected in pairs, the differences computed for each pair are the data that are used in the formulas.
- A confidence interval for the difference in two population means using paired sampling is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population mean is used to test hypotheses concerning the difference between two population means using pair sampling. The only difference is in the formula for the standardized test statistic.

This page titled 9.3: Two Population Means - Paired Samples is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **9.3: Comparison of Two Population Means - Paired Samples** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.4: Sample Size Considerations

Learning Objectives

- To learn how to apply formulas for estimating the size samples that will be needed in order to construct a confidence interval for the difference in two population means or proportions that meets given criteria.

As was pointed out at the beginning of Section 7.4, sampling is typically done with definite objectives in mind. For example, a physician might wish to estimate the difference in the average amount of sleep gotten by patients suffering a certain condition with the average amount of sleep got by healthy adults, at 90% confidence and to within half an hour. Since sampling costs time, effort, and money, it would be useful to be able to estimate the smallest size samples that are likely to meet these criteria.

Estimating $\mu_1 - \mu_2$ with Independent Samples

Assuming that large samples will be required, the confidence interval formula for estimating the difference $\mu_1 - \mu_2$ between two population means using independent samples is $(\bar{x_1} - \bar{x_2}) \pm E$, where

$$E = z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

To say that we wish to estimate the mean to within a certain number of units means that we want the margin of error E to be no larger than that number. The number $z_{\alpha/2}$ is determined by the desired level of confidence.

The numbers s_1 and s_2 are estimates of the standard deviations σ_1 and σ_2 of the two populations. In analogy with what we did in Section 7.4 we will assume that we either know or can reasonably approximate σ_1 and σ_2.

We cannot solve for both n_1 and n_2, so we have to make an assumption about their relative sizes. We will specify that they be equal. With these assumptions we obtain the minimum sample sizes needed by solving the equation displayed just above for $n_1 = n_2$.

Minimum Equal Sample Sizes for Estimating the Difference in the Means of Two Populations Using Independent Samples

The estimated minimum equal sample sizes $n_1 = n_2$ needed to estimate the difference $\mu_1 - \mu_2$ in two population means to within E units at $100(1-\alpha)\%$ confidence is

$$n_1 = n_2 = \frac{(z_{\alpha/2})^2(\sigma_1^2 + \sigma_2^2)}{E^2} \text{ rounded up}$$

In all the examples and exercises the population standard deviations σ_1 and σ_2 will be given.

Example 9.4.1

A law firm wishes to estimate the difference in the mean delivery time of documents sent between two of its offices by two different courier companies, to within half an hour and with 99.5% confidence. From their records it will randomly sample the same number n of documents as delivered by each courier company. Determine how large n must be if the estimated standard deviations of the delivery times are 0.75 hour for one company and 1.15 hours for the other.

Solution

Confidence level 99.5% means that $\alpha = 1 - 0.995 = 0.005$ so $\alpha/2 = 0.0025$. From the last line of Figure 7.1.6 we obtain $z_{0.0025} = 2.807$.

To say that the estimate is to be "to within half an hour" means that $E = 0.5$. Thus

$$n = \frac{(z_{\alpha/2})^2(\sigma_1^2 + \sigma_2^2)}{E^2} = \frac{(2.807)^2(0.75^2 + 1.15^2)}{0.5^2} = 59.40953746$$

which we round up to 60, since it is impossible to take a fractional observation. The law firm must sample 60 document deliveries by each company.

Estimating $\mu_1 - \mu_2$ with Paired Samples

As we mentioned at the end of Section 9.3, if the sample is large (meaning that $n \geq 30$) then in the formula for the confidence interval we may replace $t_{\alpha/2}$ by $z_{\alpha/2}$, so that the confidence interval formula becomes $\bar{d} \pm E$ for

$$E = z_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

The number s_d is an estimate of the standard deviations σ_d of the population of differences. We must assume that we either know or can reasonably approximate σ_d. Thus, assuming that large samples will be required to meet the criteria given, we can solve the displayed equation for n to obtain an estimate of the number of pairs needed in the sample.

Minimum Sample Size for Estimating the Difference in the Means of Two Populations Using Paired Difference Samples

The estimated minimum number of pairs n needed to estimate the difference $\mu_d = \mu_1 - \mu_2$ in two population means to within E units at $100(1-\alpha)\%$ confidence using paired difference samples is

$$n = \frac{(z_{\alpha/2})^2\sigma_d^2}{E^2} \text{ rounded up}$$

In all the examples and exercises the population standard deviation of the differences σ_d will be given.

Example 9.4.2

A automotive tire manufacturer wishes to compare the mean lifetime of two tread designs under actual driving conditions. They will mount one of each type of tire on n vehicles (both on the front or both on the back) and measure the difference in remaining tread after $20,000$ miles of driving. If the standard deviation of the differences is assumed to be 0.025 inch, find the minimum samples size needed to estimate the difference in mean depth (at $20,000$ miles use) to within 0.01 inch at 99.9% confidence.

Solution

Confidence level 99.9% means that $\alpha = 1 - 0.999 = 0.001$ so $\alpha/2 = 0.0005$. From the last line of Figure 7.1.6 we obtain $z_{0.0005} = 3.291$.

To say that the estimate is to be "to within 0.01 inch" means that $E = 0.01$. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma_d^2}{E^2} = \frac{(3.291)^2(0.025)^2}{(0.01)^2} = 67.69175625$$

which we round up to 68. The manufacturer must test 68 pairs of tires.

Estimating $p_1 - p_2$

The confidence interval formula for estimating the difference $p_1 - p_2$ between two population proportions is $\hat{p_1} - \hat{p_2} \pm E$, where

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

To say that we wish to estimate the mean to within a certain number of units means that we want the margin of error E to be no larger than that number. The number $z_{\alpha/2}$ is determined by the desired level of confidence.

We cannot solve for both n_1 and n_2, so we have to make an assumption about their relative sizes. We will specify that they be equal. With these assumptions we obtain the minimum sample sizes needed by solving the displayed equation for $n_1 = n_2$.

Minimum Equal Sample Sizes for Estimating the Difference in Two Population Proportions

The estimated minimum equal sample sizes $n_1 = n_2$ needed to estimate the difference $p_1 - p_2$ in two population proportions to within E percentage points at $100(1-\alpha)\%$ confidence is

$$n_1 = n_2 = \frac{(z_{\alpha/2})^2(\hat{p}_1(1-\hat{p}_1 + \hat{p}_2(1-\hat{p}_2))}{E^2} \text{ rounded up}$$

Here we face the same dilemma that we encountered in the case of a single population proportion: the formula for estimating how large a sample to take contains the numbers $\hat{p}_1$ and $\hat{p}_2$, which we know only after we have taken the sample. There are two ways out of this dilemma. Typically the researcher will have some idea as to the values of the population proportions p_1 and p_2, hence of what the sample proportions $\hat{p}_1$ and $\hat{p}_2$ are likely to be. If so, those estimates can be used in the formula.

The second approach to resolving the dilemma is simply to replace each of $\hat{p}_1$ and $\hat{p}_2$ in the formula by 0.5. As in the one-population case, this is the most conservative estimate, since it gives the largest possible estimate of n. If we have an estimate of only one of p_1 and p_2 we can use that estimate for it, and use the conservative estimate 0.5 for the other.

Example 9.4.3

Find the minimum equal sample sizes necessary to construct a 98% confidence interval for the difference $p_1 - p_2$ with a margin of error $E = 0.05$,

1. assuming that no prior knowledge about p_1 or p_2 is available; and
2. assuming that prior studies suggest that $p_1 \approx 0.2$ and $p_2 \approx 0.3$.

Solution

Confidence level 98% means that $\alpha = 1 - 0.98 = 0.02$ so $\alpha/2 = 0.01$. From the last line of Figure 7.1.6 we obtain $z_{0.01} = 2.326$.

1. Since there is no prior knowledge of p_1 or p_2 we make the most conservative estimate that $\hat{p}_1 = 0.5$ and $\hat{p}_2 = 0.5$. Then

$$\begin{aligned} n_1 = n_2 &= \frac{(z_{\alpha/2})^2(\hat{p}_1(1-\hat{p}_1 + \hat{p}_2(1-\hat{p}_2))}{E^2} \\ &= \frac{(2.326)^2((0.5)(0.5) + (0.5)(0.5))}{0.05^2} \\ &= 1082.0552 \end{aligned}$$

which we round up to $1,083$. We must take a sample of size $1,083$ from each population.

2. Since $p_1 \approx 0.2$ we estimate $\hat{p}_1$ by 0.2, and since $p_2 \approx 0.3$ we estimate $\hat{p}_2$ by 0.3. Thus we obtain

$$\begin{aligned} n_1 = n_2 &= \frac{(z_{\alpha/2})^2(\hat{p}_1(1-\hat{p}_1 + \hat{p}_2(1-\hat{p}_2))}{E^2} \\ &= \frac{(2.326)^2((0.2)(0.8) + (0.3)(0.7))}{0.05^2} \\ &= 800.720848 \end{aligned}$$

which we round up to 801. We must take a sample of size 801 from each population.

Key Takeaway

- If the population standard deviations σ_1 and σ_2 are known or can be estimated, then the minimum equal sizes of independent samples needed to obtain a confidence interval for the difference $\mu_1 - \mu_2$ in two population means with a given maximum error of the estimate E and a given level of confidence can be estimated.
- If the standard deviation σ_d of the population of differences in pairs drawn from two populations is known or can be estimated, then the minimum number of sample pairs needed under paired difference sampling to obtain a confidence interval for the difference $\mu_d = \mu_1 - \mu_2$ in two population means with a given maximum error of the estimate E and a given level of confidence can be estimated.

- The minimum equal sample sizes needed to obtain a confidence interval for the difference in two population proportions with a given maximum error of the estimate and a given level of confidence can always be estimated. If there is prior knowledge of the population proportions p_1 and p_2 then the estimate can be sharpened.

This page titled 9.4: Sample Size Considerations is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **9.5: Sample Size Considerations** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.E: Two-Sample Problems (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

9.1: Comparison of Two Population Means: Large, Independent Samples

Basic

Q9.1.1

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 90% confidence,

$$\begin{aligned} n_1 = 45, \bar{x}_1 = 27, s_1 = 2 \\ n_2 = 60, \bar{x}_2 = 22, s_2 = 3 \end{aligned} \tag{9.E.1}$$

b. 99% confidence,

$$\begin{aligned} n_1 = 30, \bar{x}_1 = -112, s_1 = 9 \\ n_2 = 40, \bar{x}_2 = -98, s_2 = 4 \end{aligned} \tag{9.E.2}$$

Q9.1.2

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 95% confidence,

$$\begin{aligned} n_1 = 110, \bar{x}_1 = 77, s_1 = 15 \\ n_2 = 85, \bar{x}_2 = 79, s_2 = 21 \end{aligned} \tag{9.E.3}$$

b. 90% confidence,

$$\begin{aligned} n_1 = 65, \bar{x}_1 = -83, s_1 = 12 \\ n_2 = 65, \bar{x}_2 = -74, s_2 = 8 \end{aligned} \tag{9.E.4}$$

Q9.1.3

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.5% confidence,

$$\begin{aligned} n_1 = 130, \bar{x}_1 = 27.2, s_1 = 2.5 \\ n_2 = 155, \bar{x}_2 = 38.8, s_2 = 4.6 \end{aligned} \tag{9.E.5}$$

b. 95% confidence,

$$\begin{aligned} n_1 = 68, \bar{x}_1 = 215.5, s_1 = 12.3 \\ n_2 = 84, \bar{x}_2 = 287.8, s_2 = 14.1 \end{aligned} \tag{9.E.6}$$

Q9.1.4

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.9% confidence,

$$\begin{aligned} n_1 = 275, \bar{x}_1 = 70.2, s_1 = 1.5 \\ n_2 = 325, \bar{x}_2 = 63.4, s_2 = 1.1 \end{aligned} \tag{9.E.7}$$

b. 90% confidence,

$$\begin{aligned} n_1 = 120, \bar{x}_1 = 35.5, s_1 = 0.75 \\ n_2 = 146, \bar{x}_2 = 29.6, s_2 = 0.80 \end{aligned} \tag{9.E.8}$$

Q9.1.5

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 3$ *vs* $H_a : \mu_1 - \mu_2 \neq 3$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 = 35, \bar{x_1} = 25, s_1 = 1 \\ n_2 = 45, \bar{x_2} = 19, s_2 = 2 \end{aligned} \tag{9.E.9}$$

b. Test $H_0 : \mu_1 - \mu_2 = -25$ *vs* $H_a : \mu_1 - \mu_2 < -25$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 85, \bar{x_1} = 188, s_1 = 15 \\ n_2 = 62, \bar{x_2} = 215, s_2 = 19 \end{aligned} \tag{9.E.10}$$

Q9.1.6

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 45$ *vs* $H_a : \mu_1 - \mu_2 > 45$ @ $\alpha = 0.001$

$$\begin{aligned} n_1 = 200, \bar{x_1} = 1312, s_1 = 35 \\ n_2 = 225, \bar{x_2} = 1256, s_2 = 28 \end{aligned} \tag{9.E.11}$$

b. Test $H_0 : \mu_1 - \mu_2 = -12$ *vs* $H_a : \mu_1 - \mu_2 \neq -12$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 35, \bar{x_1} = 121, s_1 = 6 \\ n_2 = 40, \bar{x_2} = 135 s_2 = 7 \end{aligned} \tag{9.E.12}$$

Q9.1.7

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 0$ *vs* $H_a : \mu_1 - \mu_2 \neq 0$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 125, \bar{x_1} = -46, s_1 = 10 \\ n_2 = 90, \bar{x_2} = -50, s_2 = 13 \end{aligned} \tag{9.E.13}$$

b. Test $H_0 : \mu_1 - \mu_2 = 20$ *vs* $H_a : \mu_1 - \mu_2 > 20$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 = 40, \bar{x_1} = 142, s_1 = 11 \\ n_2 = 40, \bar{x_2} = 118 s_2 = 10 \end{aligned} \tag{9.E.14}$$

Q9.1.8

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 13$ *vs* $H_a : \mu_1 - \mu_2 < 13$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 35, \bar{x_1} = 100, s_1 = 2 \\ n_2 = 35, \bar{x_2} = 88, s_2 = 2 \end{aligned} \tag{9.E.15}$$

b. Test $H_0 : \mu_1 - \mu_2 = -10$ *vs* $H_a : \mu_1 - \mu_2 \neq -10$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 146, \bar{x_1} = 62, s_1 = 4 \\ n_2 = 120, \bar{x_2} = 73 s_2 = 7 \end{aligned} \tag{9.E.16}$$

Q9.1.9

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 57$ *vs* $H_a : \mu_1 - \mu_2 < 57$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 117, \bar{x_1} = 1309, s_1 = 42 \\ n_2 = 133, \bar{x_2} = 1258, s_2 = 37 \end{aligned} \tag{9.E.17}$$

b. Test $H_0 : \mu_1 - \mu_2 = -1.5$ *vs* $H_a : \mu_1 - \mu_2 \neq -1.5$ @ $\alpha = 0.20$

$$\begin{aligned} n_1 = 65, \bar{x_1} = 16.9, s_1 = 1.3 \\ n_2 = 57, \bar{x_2} = 18.6 s_2 = 1.1 \end{aligned} \tag{9.E.18}$$

Q9.1.10

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = -10.5$ *vs* $H_a : \mu_1 - \mu_2 > -10.5$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 64, \bar{x_1} = 85.6, s_1 = 2.4 \\ n_2 = 50, \bar{x_2} = 95.3, s_2 = 3.1 \end{aligned} \tag{9.E.19}$$

b. Test $H_0 : \mu_1 - \mu_2 = 110$ *vs* $H_a : \mu_1 - \mu_2 \neq 110$ @ $\alpha = 0.02$

$$\begin{aligned} n_1 = 176, \bar{x_1} = 1918, s_1 = 68 \\ n_2 = 241, \bar{x_2} = 1782 s_2 = 146 \end{aligned} \tag{9.E.20}$$

Q9.1.11

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 50$ *vs* $H_a : \mu_1 - \mu_2 > 50$ @ $\alpha = 0.005$

$$\begin{aligned} n_1 = 72, \bar{x_1} = 272, s_1 = 26 \\ n_2 = 103, \bar{x_2} = 213, s_2 = 14 \end{aligned} \tag{9.E.21}$$

b. Test $H_0 : \mu_1 - \mu_2 = 7.5$ *vs* $H_a : \mu_1 - \mu_2 \neq 7.5$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 52, \bar{x_1} = 94.3, s_1 = 2.6 \\ n_2 = 38, \bar{x_2} = 88.6 s_2 = 8.0 \end{aligned} \tag{9.E.22}$$

Q9.1.12

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 23$ *vs* $H_a : \mu_1 - \mu_2 < 23$ @ $\alpha = 0.20$

$$\begin{aligned} n_1 = 314, \bar{x_1} = 198, s_1 = 12.2 \\ n_2 = 220, \bar{x_2} = 176, s_2 = 11.5 \end{aligned} \tag{9.E.23}$$

b. Test $H_0 : \mu_1 - \mu_2 = 4.4$ *vs* $H_a : \mu_1 - \mu_2 \neq 4.4$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 = 32, \bar{x_1} = 40.3, s_1 = 0.5 \\ n_2 = 30, \bar{x_2} = 35.5 s_2 = 0.7 \end{aligned} \tag{9.E.24}$$

Applications

Q9.1.13

In order to investigate the relationship between mean job tenure in years among workers who have a bachelor's degree or higher and those who do not, random samples of each type of worker were taken, with the following results.

	n	$\bar{x}$	s
Bachelor's degree or higher	155	5.2	1.3
No degree	210	5.0	1.5

a. Construct the 99% confidence interval for the difference in the population means based on these data.
b. Test, at the 1% level of significance, the claim that mean job tenure among those with higher education is greater than among those without, against the default that there is no difference in the means.
c. Compute the observed significance of the test.

Q9.1.14

Records of 40 used passenger cars and 40 used pickup trucks (none used commercially) were randomly selected to investigate whether there was any difference in the mean time in years that they were kept by the original owner before being sold. For cars the mean was 5.3 years with standard deviation 2.2 years. For pickup trucks the mean was 7.1 years with standard deviation 3.0 years.

a. Construct the 95% confidence interval for the difference in the means based on these data.
b. Test the hypothesis that there is a difference in the means against the null hypothesis that there is no difference. Use the 1% level of significance.
c. Compute the observed significance of the test in part (b).

Q9.1.15

In previous years the average number of patients per hour at a hospital emergency room on weekends exceeded the average on weekdays by 6.3 visits per hour. A hospital administrator believes that the current weekend mean exceeds the weekday mean by fewer than 6.3 hours.

a. Construct the 99% confidence interval for the difference in the population means based on the following data, derived from a study in which 30 weekend and 30 weekday one-hour periods were randomly selected and the number of new patients in each recorded.

	n	$\bar{x}$	s
Weekends	30	13.8	3.1
Weekdays	30	8.6	2.7

a. Test at the 5% level of significance whether the current weekend mean exceeds the weekday mean by fewer than 6.3 patients per hour.
b. Compute the observed significance of the test.

Q9.1.16

A sociologist surveys 50 randomly selected citizens in each of two countries to compare the mean number of hours of volunteer work done by adults in each. Among the 50 inhabitants of Lilliput, the mean hours of volunteer work per year was 52, with standard deviation 11.8. Among the 50 inhabitants of Blefuscu, the mean number of hours of volunteer work per year was 37, with standard deviation 7.2.

a. Construct the 99% confidence interval for the difference in mean number of hours volunteered by all residents of Lilliput and the mean number of hours volunteered by all residents of Blefuscu.
b. Test, at the 1% level of significance, the claim that the mean number of hours volunteered by all residents of Lilliput is more than ten hours greater than the mean number of hours volunteered by all residents of Blefuscu.
c. Compute the observed significance of the test in part (b).

Q9.1.17

A university administrator asserted that upperclassmen spend more time studying than underclassmen.

a. Test this claim against the default that the average number of hours of study per week by the two groups is the same, using the following information based on random samples from each group of students. Test at the 1% level of significance.

	n	$\bar{x}$	s
Upperclassmen	35	15.6	2.9
Underclassmen	35	12.3	4.1

b. Compute the observed significance of the test.

Q9.1.18

An kinesiologist claims that the resting heart rate of men aged 18 to 25 who exercise regularly is more than five beats per minute less than that of men who do not exercise regularly. Men in each category were selected at random and their resting heart rates were measured, with the results shown.

	n	$\bar{x}$	s
Regular exercise	40	63	1.0
No regular exercise	30	71	1.2

a. Perform the relevant test of hypotheses at the 1% level of significance.
b. Compute the observed significance of the test.

Q9.1.19

Children in two elementary school classrooms were given two versions of the same test, but with the order of questions arranged from easier to more difficult in Version A and in reverse order in Version B. Randomly selected students from each class were given Version A and the rest Version B. The results are shown in the table.

	n	$\bar{x}$	s
Version A	31	83	4.6
Version B	32	78	4.3

a. Construct the 90% confidence interval for the difference in the means of the populations of all children taking Version A of such a test and of all children taking Version B of such a test.
b. Test at the 1% level of significance the hypothesis that the A version of the test is easier than the B version (even though the questions are the same).
c. Compute the observed significance of the test.

Q9.1.20

The Municipal Transit Authority wants to know if, on weekdays, more passengers ride the northbound blue line train towards the city center that departs at $8:15\ a.m.$ or the one that departs at $8:30\ a.m.$. The following sample statistics are assembled by the Transit Authority.

	n	$\bar{x}$	s
8:15 a.m. train	30	323	41
8:30 a.m. train	45	356	45

a. Construct the 90% confidence interval for the difference in the mean number of daily travelers on the $8:15\ a.m.$ train and the mean number of daily travelers on the $8:30\ a.m.$ train.
b. Test at the 5% level of significance whether the data provide sufficient evidence to conclude that more passengers ride the $8:30\ a.m.$ train.
c. Compute the observed significance of the test.

Q9.1.21

In comparing the academic performance of college students who are affiliated with fraternities and those male students who are unaffiliated, a random sample of students was drawn from each of the two populations on a university campus. Summary statistics on the student GPAs are given below.

	n	$\bar{x}$	s
Fraternity	645	2.90	0.47
Unaffiliated	450	2.88	0.42

Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that there is a difference in average GPA between the population of fraternity students and the population of unaffiliated male students on this university campus.

Q9.1.22

In comparing the academic performance of college students who are affiliated with sororities and those female students who are unaffiliated, a random sample of students was drawn from each of the two populations on a university campus. Summary statistics on the student GPAs are given below.

	n	$\bar{x}$	s
Sorority	330	3.18	0.37
Unaffiliated	550	3.12	0.41

Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that there is a difference in average GPA between the population of sorority students and the population of unaffiliated female students on this university campus.

Q9.1.23

The owner of a professional football team believes that the league has become more offense oriented since five years ago. To check his belief, 32 randomly selected games from one year's schedule were compared to 32 randomly selected games from the schedule five years later. Since more offense produces more points per game, the owner analyzed the following information on points per game (ppg).

	n	$\bar{x}$	s
ppg previously	32	20.62	4.17
ppg recently	32	22.05	4.01

Test, at the 10% level of significance, whether the data on points per game provide sufficient evidence to conclude that the game has become more offense oriented.

Q9.1.24

The owner of a professional football team believes that the league has become more offense oriented since five years ago. To check his belief, 32 randomly selected games from one year's schedule were compared to 32 randomly selected games from the schedule five years later. Since more offense produces more offensive yards per game, the owner analyzed the following information on offensive yards per game (oypg).

	n	$\bar{x}$	s
oypg previously	32	316	40
oypg recently	32	336	35

Test, at the 10% level of significance, whether the data on offensive yards per game provide sufficient evidence to conclude that the game has become more offense oriented.

Large Data Set Exercises

Large Data Sets are absent

25. Large **Data Sets 1A and 1B** list the SAT scores for 1,000 randomly selected students. Denote the population of all male students as **Population 1** and the population of all female students as **Population 2**.
    a. Restricting attention to just the males, find n_1, $\bar{x}_1$ and s_1. Restricting attention to just the females, find n_2, $\bar{x}_2$ and s_2.
    b. Let μ_1 denote the mean SAT score for all males and μ_2 the mean SAT score for all females. Use the results of part (a) to construct a 90% confidence interval for the difference $\mu_1 - \mu_2$.

c. Test, at the 5% level of significance, the hypothesis that the mean SAT scores among males exceeds that of females.

26. Large **Data Sets 1A and 1B** list the SAT scores for 1,000 randomly selected students. Denote the population of all male students as **Population 1** and the population of all female students as **Population 2**.
    a. Restricting attention to just the males, find n_1, $\bar{x}_1$ and s_1. Restricting attention to just the females, find n_2, $\bar{x}_2$ and s_2.
    b. Let μ_1 denote the mean SAT score for all males and μ_2 the mean SAT score for all females. Use the results of part (a) to construct a 95% confidence interval for the difference $\mu_1 - \mu_2$.
    c. Test, at the 10% level of significance, the hypothesis that the mean SAT scores among males exceeds that of females.
27. Large **Data Sets 7A and 7B** list the survival times for 65 male and 75 female laboratory mice with thymic leukemia. Denote the population of all such male mice as **Population 1** and the population of all such female mice as **Population 2**.
    a. Restricting attention to just the males, find n_1, $\bar{x}_1$ and s_1. Restricting attention to just the females, find n_2, $\bar{x}_2$ and s_2.
    b. Let μ_1 denote the mean survival for all males and μ_2 the mean survival time for all females. Use the results of part (a) to construct a 99% confidence interval for the difference $\mu_1 - \mu_2$.
    c. Test, at the 1% level of significance, the hypothesis that the mean survival time for males exceeds that for females by more than 182 days (half a year).
    d. Compute the observed significance of the test in part (c).

Answers

1. a. $(4.20, 5.80)$
   b. $(-18.54, -9.46)$
2.
3. a. $(-12.81, -10.39)$
   b. $(-76.50, -68.10)$
4.
5. a. $Z = 8.753, \pm z_{0.025} = \pm 1.960$, reject H_0, p-value=0.0000
   b. $Z = -0.687, -z_{0.10} = -1.282$, do not reject H_0, p-value=0.2451
6.
7. a. $Z = 2.444, \pm z_{0.005} = \pm 2.576$, do not reject H_0, p-value=0.0146
   b. $Z = 1.702, z_{0.05} = -1.645$, reject H_0, p-value=0.0446
8.
9. a. $Z = -1.19$, p-value=0.1170, do not reject H_0
   b. $Z = -0.92$, p-value=0.3576, do not reject H_0
10.
11. a. $Z = 2.68$, p-value=0.0037, reject H_0
    b. $Z = -1.34$, p-value=0.1802, do not reject H_0
12.
13. a. 0.2 ± 0.4
    b. $Z = -1.466, -z_{0.050} = -1.645$, do not reject H_0 (exceeds by 6.3 or more)
    c. p-value=0.0869
14.
15. a. 5.2 ± 1.9
    b. $Z = -1.466, -z_{0.050} = -1.645$, do not reject H_0 (exceeds by 6.3 or more)
    c. p-value=0.0708
16.
17. a. $Z = 3.888, z_{0.01} = 2.326$, reject H_0 (upperclassmen study more)
    b. p-value=0.0001
18.

19. a. 5 ± 1.8
    b. $Z = 4.454, z_{0.01} = 2.326$, reject H_0 (Test A is easier)
    c. p-value=0.0000
20.
21. $Z = 0.738, \pm z_{0.025} = \pm 1.960$, do not reject H_0 (no difference)
22.
23. $Z = -1.398, -z_{0.10} = -1.282$, reject H_0 (more offense oriented)
24.
25. a. $n_1 = 419,\ \bar{x}_1 = 1540.33,\ s_1 = 205.40,\ n_2 = 581,\ \bar{x}_2 = 1520.38,\ s_2 = 217.34$
    b. $(-2.24, 42.15)$
    c. $H_0 : \mu_1 - \mu_2 = 0\ vs\ H_a : \mu_1 - \mu_2 > 0$. Test Statistic: $Z = 1.48$. Rejection Region: $[1.645, \infty)$. Decision: Fail to reject H_0.
26.
27. a. $n_1 = 65,\ \bar{x}_1 = 665.97,\ s_1 = 41.60,\ n_2 = 75,\ \bar{x}_2 = 455.89,\ s_2 = 63.22$
    b. $(187.06, 233.09)$
    c. $H_0 : \mu_1 - \mu_2 = 182\ vs\ H_a : \mu_1 - \mu_2 > 182$. Test Statistic: $Z = 3.14$. Rejection Region: $[2.33, \infty)$. Decision: Reject H_0.
    d. p-value=0.0008

9.2: Comparison of Two Population Means: Small, Independent Samples

Basic

In all exercises for this section assume that the populations are normal and have equal standard deviations.

Q9.2.1

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 95% confidence,

$$\begin{aligned} n_1 &= 10,\ \bar{x}_1 = 120,\ s_1 = 2 \\ n_2 &= 15,\ \bar{x}_2 = 101,\ s_1 = 4 \end{aligned} \tag{9.E.25}$$

b. 99% confidence,

$$\begin{aligned} n_1 &= 6,\ \bar{x}_1 = 25,\ s_1 = 1 \\ n_2 &= 12,\ \bar{x}_2 = 17,\ s_1 = 3 \end{aligned} \tag{9.E.26}$$

Q9.2.2

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 90% confidence,

$$\begin{aligned} n_1 &= 28,\ \bar{x}_1 = 212,\ s_1 = 6 \\ n_2 &= 23,\ \bar{x}_2 = 198,\ s_1 = 5 \end{aligned} \tag{9.E.27}$$

b. 99% confidence,

$$\begin{aligned} n_1 &= 14,\ \bar{x}_1 = 68,\ s_1 = 8 \\ n_2 &= 20,\ \bar{x}_2 = 43,\ s_1 = 3 \end{aligned} \tag{9.E.28}$$

Q9.2.3

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.9% confidence,

$$\begin{aligned} n_1 &= 35,\ \bar{x}_1 = 6.5,\ s_1 = 0.2 \\ n_2 &= 20,\ \bar{x}_2 = 6.2,\ s_1 = 0.1 \end{aligned} \tag{9.E.29}$$

b. 99% confidence,

$$n_1 = 18,\ \bar{x_1} = 77.3,\ s_1 = 1.2 \\ n_2 = 32,\ \bar{x_2} = 75.0,\ s_1 = 1.6 \tag{9.E.30}$$

Q9.2.4

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.5% confidence,

$$n_1 = 40,\ \bar{x_1} = 85.6,\ s_1 = 2.8 \\ n_2 = 20,\ \bar{x_2} = 73.1,\ s_1 = 2.1 \tag{9.E.31}$$

b. 99.9% confidence,

$$n_1 = 25,\ \bar{x_1} = 215,\ s_1 = 7 \\ n_2 = 35,\ \bar{x_2} = 185,\ s_1 = 12 \tag{9.E.32}$$

Q9.2.5

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 11$ *vs* $H_a : \mu_1 - \mu_2 > 11$ @ $\alpha = 0.025$

$$n_1 = 6,\ \bar{x_1} = 32,\ s_1 = 2 \\ n_2 = 11,\ \bar{x_2} = 19,\ s_1 = 1 \tag{9.E.33}$$

b. Test $H_0 : \mu_1 - \mu_2 = 26$ *vs* $H_a : \mu_1 - \mu_2 \neq 26$ @ $\alpha = 0.05$

$$n_1 = 17,\ \bar{x_1} = 166,\ s_1 = 4 \\ n_2 = 24,\ \bar{x_2} = 138,\ s_1 = 3 \tag{9.E.34}$$

Q9.2.6

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 40$ *vs* $H_a : \mu_1 - \mu_2 < 40$ @ $\alpha = 0.10$

$$n_1 = 14,\ \bar{x_1} = 289,\ s_1 = 11 \\ n_2 = 12,\ \bar{x_2} = 254,\ s_1 = 9 \tag{9.E.35}$$

b. Test $H_0 : \mu_1 - \mu_2 = 21$ *vs* $H_a : \mu_1 - \mu_2 \neq 21$ @ $\alpha = 0.05$

$$n_1 = 23,\ \bar{x_1} = 130,\ s_1 = 6 \\ n_2 = 27,\ \bar{x_2} = 113,\ s_1 = 8 \tag{9.E.36}$$

Q9.2.7

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = -15$ *vs* $H_a : \mu_1 - \mu_2 < -15$ @ $\alpha = 0.10$

$$n_1 = 30,\ \bar{x_1} = 42,\ s_1 = 7 \\ n_2 = 12,\ \bar{x_2} = 60,\ s_1 = 5 \tag{9.E.37}$$

b. Test $H_0 : \mu_1 - \mu_2 = 103$ *vs* $H_a : \mu_1 - \mu_2 \neq 103$ @ $\alpha = 0.10$

$$n_1 = 17,\ \bar{x_1} = 711,\ s_1 = 28 \\ n_2 = 32,\ \bar{x_2} = 598,\ s_1 = 21 \tag{9.E.38}$$

Q9.2.8

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 75$ *vs* $H_a : \mu_1 - \mu_2 > 75$ @ $\alpha = 0.025$

$$n_1 = 45,\ \bar{x_1} = 674,\ s_1 = 18 \\ n_2 = 29,\ \bar{x_2} = 591,\ s_1 = 13 \tag{9.E.39}$$

b. Test $H_0 : \mu_1 - \mu_2 = -20$ *vs* $H_a : \mu_1 - \mu_2 \neq -20$ @ $\alpha = 0.005$

$$n_1 = 30,\ \bar{x}_1 = 137,\ s_1 = 8 \\ n_2 = 19,\ \bar{x}_2 = 166,\ s_1 = 11 \tag{9.E.40}$$

Q9.2.9

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = 12$ *vs* $H_a : \mu_1 - \mu_2 > 12$ @ $\alpha = 0.01$

$$n_1 = 20,\ \bar{x}_1 = 133,\ s_1 = 7 \\ n_2 = 10,\ \bar{x}_2 = 115,\ s_1 = 5 \tag{9.E.41}$$

b. Test $H_0 : \mu_1 - \mu_2 = 46$ *vs* $H_a : \mu_1 - \mu_2 \neq 46$ @ $\alpha = 0.10$

$$n_1 = 24,\ \bar{x}_1 = 586,\ s_1 = 11 \\ n_2 = 27,\ \bar{x}_2 = 535,\ s_1 = 13 \tag{9.E.42}$$

Q9.2.10

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = 38$ *vs* $H_a : \mu_1 - \mu_2 < 38$ @ $\alpha = 0.01$

$$n_1 = 12,\ \bar{x}_1 = 464,\ s_1 = 5 \\ n_2 = 10,\ \bar{x}_2 = 432,\ s_1 = 6 \tag{9.E.43}$$

b. Test $H_0 : \mu_1 - \mu_2 = 4$ *vs* $H_a : \mu_1 - \mu_2 \neq 4$ @ $\alpha = 0.005$

$$n_1 = 14,\ \bar{x}_1 = 68,\ s_1 = 2 \\ n_2 = 17,\ \bar{x}_2 = 67,\ s_1 = 3 \tag{9.E.44}$$

Q9.2.11

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = 50$ *vs* $H_a : \mu_1 - \mu_2 > 50$ @ $\alpha = 0.01$

$$n_1 = 30,\ \bar{x}_1 = 681,\ s_1 = 8 \\ n_2 = 27,\ \bar{x}_2 = 625,\ s_1 = 8 \tag{9.E.45}$$

b. Test $H_0 : \mu_1 - \mu_2 = 35$ *vs* $H_a : \mu_1 - \mu_2 \neq 35$ @ $\alpha = 0.10$

$$n_1 = 36,\ \bar{x}_1 = 325,\ s_1 = 11 \\ n_2 = 29,\ \bar{x}_2 = 286,\ s_1 = 7 \tag{9.E.46}$$

Q9.2.12

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = -4$ *vs* $H_a : \mu_1 - \mu_2 < -4$ @ $\alpha = 0.05$

$$n_1 = 40,\ \bar{x}_1 = 80,\ s_1 = 5 \\ n_2 = 25,\ \bar{x}_2 = 87,\ s_1 = 5 \tag{9.E.47}$$

b. Test $H_0 : \mu_1 - \mu_2 = 21$ *vs* $H_a : \mu_1 - \mu_2 \neq 21$ @ $\alpha = 0.01$

$$n_1 = 15,\ \bar{x}_1 = 192,\ s_1 = 12 \\ n_2 = 34,\ \bar{x}_2 = 180,\ s_1 = 8 \tag{9.E.48}$$

Applications

Q9.2.13

A county environmental agency suspects that the fish in a particular polluted lake have elevated mercury level. To confirm that suspicion, five striped bass in that lake were caught and their tissues were tested for mercury. For the purpose of comparison, four striped bass in an unpolluted lake were also caught and tested. The fish tissue mercury levels in mg/kg are given below.

Sample 1 (from polluted lake)	Sample 2(from unpolluted lake)
0.580	0.382
0.711	0.276
0.571	0.570
0.666	0.366
0.598	

a. Construct the 95% confidence interval for the difference in the population means based on these data.
b. Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that fish in the polluted lake have elevated levels of mercury in their tissue.

Q9.2.14

A genetic engineering company claims that it has developed a genetically modified tomato plant that yields on average more tomatoes than other varieties. A farmer wants to test the claim on a small scale before committing to a full-scale planting. Ten genetically modified tomato plants are grown from seeds along with ten other tomato plants. At the season's end, the resulting yields in pound are recorded as below.

Sample 1(genetically modified)	Sample 2(regular)
20	21
23	21
27	22
25	18
25	20
25	20
27	18
23	25
24	23
22	20

a. Construct the 99% confidence interval for the difference in the population means based on these data.
b. Test, at the 1% level of significance, whether the data provide sufficient evidence to conclude that the mean yield of the genetically modified variety is greater than that for the standard variety.

Q9.2.15

The coaching staff of a professional football team believes that the rushing offense has become increasingly potent in recent years. To investigate this belief, 20 randomly selected games from one year's schedule were compared to 11 randomly selected games from the schedule five years later. The sample information on rushing yards per game (rypg) is summarized below.

	n	$\bar{x}$	s
rypg previously	20	112	24
rypg recently	11	114	21

a. Construct the 95% confidence interval for the difference in the population means based on these data.
b. Test, at the 5% level of significance, whether the data on rushing yards per game provide sufficient evidence to conclude that the rushing offense has become more potent in recent years.

Q9.2.16

The coaching staff of professional football team believes that the rushing offense has become increasingly potent in recent years. To investigate this belief, 20 randomly selected games from one year's schedule were compared to 11 randomly selected games from the schedule five years later. The sample information on passing yards per game (pypg) is summarized below.

	n	$\bar{x}$	s
pypg previously	20	203	38
pypg recently	11	232	33

a. Construct the 95% confidence interval for the difference in the population means based on these data.
b. Test, at the 5% level of significance, whether the data on passing yards per game provide sufficient evidence to conclude that the passing offense has become more potent in recent years.

Q9.2.17

A university administrator wishes to know if there is a difference in average starting salary for graduates with master's degrees in engineering and those with master's degrees in business. Fifteen recent graduates with master's degree in engineering and 11 with master's degrees in business are surveyed and the results are summarized below.

	n	$\bar{x}$	s
Engineering	15	68,535	1627
Business	11	63,230	2033

a. Construct the 90% confidence interval for the difference in the population means based on these data.
b. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the average starting salaries are different.

Q9.2.18

A gardener sets up a flower stand in a busy business district and sells bouquets of assorted fresh flowers on weekdays. To find a more profitable pricing, she sells bouquets for 15 dollars each for ten days, then for 10 dollars each for five days. Her average daily profit for the two different prices are given below.

	n	$\bar{x}$	s
$15	10	171	26
$10	5	198	29

a. Construct the 90% confidence interval for the difference in the population means based on these data.
b. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude the gardener's average daily profit will be higher if the bouquets are sold at $10 each.

Answers

1. a. $(16.16, 21.84)$
   b. $(4.28, 11.72)$
2.
3. a. $(0.13, 0.47)$
   b. \((1.14,3.46)\)
4.
5. a. $T = 2.787,\ t_{0.025} = 2.131$, reject H_0
   b. $T = 1.831,\ \pm t_{0.025} = \pm 2.023$, do not reject H_0
6.
7. a. $T = -1.349,\ -t_{0.10} = -1.303$, reject H_0
   b. $T = 1.411,\ \pm t_{0.05} = \pm 1.678$, do not reject H_0
8.
9. a. $T = 2.411,\ df = 28,\ \text{p-value} > 0.01$, do not reject H_0
   b. $T = 1.473,\ df = 49,\ \text{p-value} < 0.10$, reject H_0
10.
11. a. $T = 2.827,\ df = 55,\ \text{p-value} < 0.01$, reject H_0
    b. $T = 1.699,\ df = 63,\ \text{p-value} < 0.10$, reject H_0
12.
13. a. 0.2267 ± 0.2182
    b. $T = 1.699,\ df = 63,\ t_{0.05} = 1.895$, reject H_0 (elevated levels)
14.
15. a. -2 ± 17.7
    b. $T = -0.232,\ df = 29,\ -t_{0.05} = -1.699$, do not reject H_0 (not more potent)
16.
17. a. 5305 ± 1227
    b. $T = 7.395,\ df = 24,\ \pm t_{0.05} = \pm 1.711$, reject H_0 (different)

9.3 Comparison of Two Population Means: Paired Samples

Basic

In all exercises for this section assume that the population of differences is normal.

1. Use the following paired sample data for this exercise.

$$\begin{array}{lccccccc} \textit{Population } 1 & 35 & 32 & 35 & 35 & 36 & 35 & 35 \\ \textit{Population } 2 & 28 & 26 & 27 & 26 & 29 & 27 & 29 \end{array} \tag{9.E.49}$$

   a. Compute $\bar{d}$ and s_d.
   b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
   c. Construct the 95% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
   d. Test, at the 10% level of significance, the hypothesis that $\mu_1 - \mu_2 > 7$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = 7$.

2. Use the following paired sample data for this exercise.

$$\begin{array}{lcccc} \textit{Population } 1 & 103 & 127 & 96 & 110 \\ \textit{Population } 2 & 81 & 106 & 73 & 88 \\ \textit{Population } 1 & 90 & 118 & 130 & 106 \\ \textit{Population } 2 & 70 & 95 & 109 & 83 \end{array} \tag{9.E.50}$$

   a. Compute $\bar{d}$ and s_d.

b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
c. Construct the 90% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
d. Test, at the 1% level of significance, the hypothesis that $\mu_1 - \mu_2 < 247$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = 24$.

3. Use the following paired sample data for this exercise.

$$\begin{array}{lccccc} \textit{Population 1} & 40 & 27 & 55 & 34 \\ \textit{Population 2} & 53 & 42 & 68 & 50 \end{array} \tag{9.E.51}$$

a. Compute $\bar{d}$ and s_d.
b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
c. Construct the 99% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
d. Test, at the 10% level of significance, the hypothesis that $\mu_1 - \mu_2 \neq -12$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = -12$.

4. Use the following paired sample data for this exercise.

$$\begin{array}{lccccc} \textit{Population 1} & 196 & 165 & 181 & 201 & 190 \\ \textit{Population 2} & 212 & 182 & 199 & 210 & 205 \end{array} \tag{9.E.52}$$

a. Compute $\bar{d}$ and s_d.
b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
c. Construct the 98% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
d. Test, at the 2% level of significance, the hypothesis that $\mu_1 - \mu_2 \neq -20$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = -20$.

Applications

5. Each of five laboratory mice was released into a maze twice. The five pairs of times to escape were:

Mouse	1	2	3	4	5
First release	129	89	136	163	118
Second release	113	97	139	85	75

a. Compute $\bar{d}$ and s_d.
b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
c. Construct the 90% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
d. Test, at the 10% level of significance, the hypothesis that it takes mice less time to run the maze on the second trial, on average.

6. Eight golfers were asked to submit their latest scores on their favorite golf courses. These golfers were each given a set of newly designed clubs. After playing with the new clubs for a few months, the golfers were again asked to submit their latest scores on the same golf courses. The results are summarized below.

Golfer	1	2	3	4	5	6	7	8
Own clubs	77	80	69	73	73	72	75	77
New clubs	72	81	68	73	75	70	73	75

a. Compute $\bar{d}$ and s_d.
b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
c. Construct the 99% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
d. Test, at the 1% level of significance, the hypothesis that on average golf scores are lower with the new clubs.

7. A neighborhood home owners association suspects that the recent appraisal values of the houses in the neighborhood conducted by the county government for taxation purposes is too high. It hired a private company to appraise the values of ten houses in

the neighborhood. The results, in thousands of dollars, are

House	County Government	Private Company
1	217	219
2	350	338
3	296	291
4	237	237
5	237	235
6	272	269
7	257	239
8	277	275
9	312	320
10	335	335

a. Give a point estimate for the difference between the mean private appraisal of all such homes and the government appraisal of all such homes.
b. Construct the 99% confidence interval based on these data for the difference.
c. Test, at the 1% level of significance, the hypothesis that appraised values by the county government of all such houses is greater than the appraised values by the private appraisal company.

8. In order to cut costs a wine producer is considering using duo or $1+1$ corks in place of full natural wood corks, but is concerned that it could affect buyers's perception of the quality of the wine. The wine producer shipped eight pairs of bottles of its best young wines to eight wine experts. Each pair includes one bottle with a natural wood cork and one with a duo cork. The experts are asked to rate the wines on a one to ten scale, higher numbers corresponding to higher quality. The results are:

Wine Expert	Duo Cork	Wood Cork
1	8.5	8.5
2	8.0	8.5
3	6.5	8.0
4	7.5	8.5
5	8.0	7.5
6	8.0	8.0
7	9.0	9.0
8	7.0	7.5

a. Give a point estimate for the difference between the mean ratings of the wine when bottled are sealed with different kinds of corks.
b. Construct the 90% confidence interval based on these data for the difference.
c. Test, at the 10% level of significance, the hypothesis that on the average duo corks decrease the rating of the wine.

9. Engineers at a tire manufacturing corporation wish to test a new tire material for increased durability. To test the tires under realistic road conditions, new front tires are mounted on each of 11 company cars, one tire made with a production material and the other with the experimental material. After a fixed period the 11 pairs were measured for wear. The amount of wear for each tire (in mm) is shown in the table:

Car	Production	Experimental
1	5.1	5.0
2	6.5	6.5
3	3.6	3.1
4	3.5	3.7
5	5.7	4.5
6	5.0	4.1
7	6.4	5.3
8	4.7	2.6
9	3.2	3.0
10	3.5	3.5
11	6.4	5.1

a. Give a point estimate for the difference in mean wear.
b. Construct the 99% confidence interval for the difference based on these data.
c. Test, at the 1% level of significance, the hypothesis that the mean wear with the experimental material is less than that for the production material.

10. A marriage counselor administered a test designed to measure overall contentment to 30 randomly selected married couples. The scores for each couple are given below. A higher number corresponds to greater contentment or happiness.

https://stats.libretexts.org/@go/page/48583

Couple	Husband	Wife
1	47	44
2	44	46
3	49	44
4	53	44
5	42	43
6	45	45
7	48	47
8	45	44
9	52	44
10	47	42
11	40	34
12	45	42
13	40	43
14	46	41
15	47	45
16	46	45
17	46	41
18	46	41
19	44	45
20	45	43
21	48	38
22	42	46
23	50	44
24	46	51
25	43	45
26	50	40
27	46	46
28	42	41
29	51	41
30	46	47

a. Test, at the 1% level of significance, the hypothesis that on average men and women are not equally happy in marriage.
b. Test, at the 1% level of significance, the hypothesis that on average men are happier than women in marriage.

Large Data Set Exercises

Large Data Sets are absent

11. Large **Data Set 5** lists the scores for **25** randomly selected students on practice SAT reading tests before and after taking a two-week SAT preparation course. Denote the population of all students who have taken the course as **Population 1** and the population of all students who have not taken the course as **Population 2**.
    a. Compute the **25** differences in the order **after - before**, their mean $\bar{d}$, and their sample standard deviation s_d.
    b. Give a point estimate for $\mu_d = \mu_1 - \mu_2$, the difference in the mean score of all students who have taken the course and the mean score of all who have not.
    c. Construct a **98%** confidence interval for μ_d.
    d. Test, at the **1%** level of significance, the hypothesis that the mean SAT score increases by at least ten points by taking the two-week preparation course.
12. Large **Data Set 12** lists the scores on one round for 75 randomly selected members at a golf course, first using their own original clubs, then two months later after using new clubs with an experimental design. Denote the population of all golfers using their own original clubs as **Population 1** and the population of all golfers using the new style clubs as **Population 2**.
    a. Compute the **75** differences in the order **original clubs - new clubs**, their mean $\bar{d}$, and their sample standard deviation s_d.
    b. Give a point estimate for $\mu_d = \mu_1 - \mu_2$, the difference in the mean score of all students who have taken the course and the mean score of all who have not.
    c. Construct a **90%** confidence interval for μ_d.
    d. Test, at the **1%** level of significance, the hypothesis that the mean SAT score increases by at least ten points by taking the two-week preparation course.
13. Consider the previous problem again. Since the data set is so large, it is reasonable to use the standard normal distribution instead of Student's t-distribution with **74** degrees of freedom.
    a. Construct a 90% confidence interval for μ_d using the standard normal distribution, meaning that the formula is $\bar{d} \pm z_{\alpha/2}\frac{s_d}{\sqrt{n}}$. (The computations done in part (a) of the previous problem still apply and need not be redone.) How does the result obtained here compare to the result obtained in part (c) of the previous problem?
    b. Test, at the **1%** level of significance, the hypothesis that the mean golf score decreases by at least one stroke by using the new kind of clubs, using the standard normal distribution. (All the work done in part (d) of the previous problem applies, except the critical value is now z_α instead of z_α (or the p-value can be computed exactly instead of only approximated, if you used the p-value approach).) How does the result obtained here compare to the result obtained in part (c) of the previous problem?
    c. Construct the **99%** confidence intervals for μ_d using both the t- and z-distributions. How much difference is there in the results now?

Answers

1. a. $\bar{d} = 7.4286,\ s_d = 0.9759$
   b. $\bar{d} = 7.4286$
   c. $(6.53, 8.33)$
   d. $T = 1.162,\ df = 6,\ t_{0.10} = 1.44$, do not reject H_0
2.
3. a. $\bar{d} = -14.25,\ s_d = 1.5$
   b. $\bar{d} = -14.25$
   c. $(-18.63, -9.87)$
   d. $T = -3.000,\ df = 3,\ \pm t_{0.05} = \pm 2.353$, reject H_0
4.
5. a. $\bar{d} = 25.2,\ s_d = 35.6609$
   b. $\bar{d} = 25.2$
   c. 25.2 ± 34.0
   d. $T = 1.580,\ df = 4,\ t_{0.10} = 1.533$, reject H_0 (takes less time)

6.

7. a. 3.2
   b. 3.2 ± 7.5
   c. $T = 1.392,\ df = 9,\ t_{0.10} = 2.821$, do not reject H_0 (government appraisals not higher)

8.

9. a. 0.65
   b. 0.65 ± 0.69
   c. $T = 3.014,\ df = 10,\ t_{0.10} = 2.764$, reject H_0 (experimental material wears less)

10.

11. a. $\bar{d} = 16.68,\ s_d = 10.77$
    b. $\bar{d} = 16.68$
    c. $(11.31, 22.05)$
    d. $H_0 : \mu_1 - \mu_2 = 10\ vs\ H_a : \mu_1 - \mu_2 > 10$. Test Statistic: $T = 3.1014,\ df = 11$. Rejection Region: $[2.492, \infty)$. Decision: Reject H_0.

12.

13. a. $(1.6266, 2.6401)$. Endpoints change in the third decimal place.
    b. $H_0 : \mu_1 - \mu_2 = 1\ vs\ H_a : \mu_1 - \mu_2 > 1$. Test Statistic: $Z = 3.6791$. Rejection Region: $[2.33, \infty)$. Decision: Reject H_0. The decision is the same as in the previous problem.
    c. Using the t-distribution, $(1.3188, 2.9478)$. Using the z-distribution, $(1.3401, 2.9266)$. There is a difference.

9.4: Comparison of Two Population Proportions

Basic

1. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

   a. 90% confidence

   $$\begin{aligned} n_1 &= 1670,\ \hat{p_1} = 0.42 \\ n_2 &= 900,\ \hat{p_2} = 0.38 \end{aligned} \tag{9.E.53}$$

   b. 95% confidence

   $$\begin{aligned} n_1 &= 600,\ \hat{p_1} = 0.84 \\ n_2 &= 420,\ \hat{p_2} = 0.67 \end{aligned} \tag{9.E.54}$$

2. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

   a. 98% confidence

   $$\begin{aligned} n_1 &= 750,\ \hat{p_1} = 0.64 \\ n_2 &= 800,\ \hat{p_2} = 0.51 \end{aligned} \tag{9.E.55}$$

   b. 99.5% confidence

   $$\begin{aligned} n_1 &= 250,\ \hat{p_1} = 0.78 \\ n_2 &= 250,\ \hat{p_2} = 0.51 \end{aligned} \tag{9.E.56}$$

3. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

   a. 80% confidence

   $$\begin{aligned} n_1 &= 300,\ \hat{p_1} = 0.255 \\ n_2 &= 400,\ \hat{p_2} = 0.193 \end{aligned} \tag{9.E.57}$$

   b. 95% confidence

   $$\begin{aligned} n_1 &= 3500,\ \hat{p_1} = 0.147 \\ n_2 &= 3750,\ \hat{p_2} = 0.131 \end{aligned} \tag{9.E.58}$$

4. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

a. 99% confidence

$$n_1 = 2250,\ \hat{p}_1 = 0.915 \\ n_2 = 2525,\ \hat{p}_2 = 0.858 \tag{9.E.59}$$

b. 95% confidence

$$n_1 = 120,\ \hat{p}_1 = 0.650 \\ n_2 = 200,\ \hat{p}_2 = 0.505 \tag{9.E.60}$$

5. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 > 0$ @ $\alpha = 0.10$

$$n_1 = 1200,\ \hat{p}_1 = 0.42 \\ n_2 = 1200,\ \hat{p}_2 = 0.40 \tag{9.E.61}$$

b. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 \neq 0$ @ $\alpha = 0.05$

$$n_1 = 550,\ \hat{p}_1 = 0.61 \\ n_2 = 600,\ \hat{p}_2 = 0.67 \tag{9.E.62}$$

6. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.05$ *vs* $H_a : p_1 - p_2 > 0.05$ @ $\alpha = 0.05$

$$n_1 = 1100,\ \hat{p}_1 = 0.57 \\ n_2 = 1100,\ \hat{p}_2 = 0.48 \tag{9.E.63}$$

b. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 \neq 0$ @ $\alpha = 0.05$

$$n_1 = 800,\ \hat{p}_1 = 0.39 \\ n_2 = 900,\ \hat{p}_2 = 0.43 \tag{9.E.64}$$

7. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.25$ *vs* $H_a : p_1 - p_2 < 0.25$ @ $\alpha = 0.005$

$$n_1 = 1400,\ \hat{p}_1 = 0.57 \\ n_2 = 1200,\ \hat{p}_2 = 0.37 \tag{9.E.65}$$

b. Test $H_0 : p_1 - p_2 = 0.16$ *vs* $H_a : p_1 - p_2 \neq 0.16$ @ $\alpha = 0.02$

$$n_1 = 750,\ \hat{p}_1 = 0.43 \\ n_2 = 600,\ \hat{p}_2 = 0.22 \tag{9.E.66}$$

8. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.08$ *vs* $H_a : p_1 - p_2 > 0.08$ @ $\alpha = 0.025$

$$n_1 = 450,\ \hat{p}_1 = 0.67 \\ n_2 = 200,\ \hat{p}_2 = 0.52 \tag{9.E.67}$$

b. Test $H_0 : p_1 - p_2 = 0.02$ *vs* $H_a : p_1 - p_2 \neq 0.02$ @ $\alpha = 0.001$

$$n_1 = 2700,\ \hat{p}_1 = 0.837 \\ n_2 = 2900,\ \hat{p}_2 = 0.854 \tag{9.E.68}$$

9. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 < 0$ @ $\alpha = 0.005$

$$n_1 = 1100,\ \hat{p}_1 = 0.22 \quad n_2 = 1300,\ \hat{p}_2 = 0.27 \tag{9.E.69}$$

b. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 \neq 0$ @ $\alpha = 0.01$

$$n_1 = 650,\ \hat{p}_1 = 0.35 \quad n_2 = 650,\ \hat{p}_2 = 0.41 \tag{9.E.70}$$

10. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.15$ *vs* $H_a : p_1 - p_2 > 0.15$ @ $\alpha = 0.10$

$$n_1 = 950,\ \hat{p}_1 = 0.41 \quad n_2 = 500,\ \hat{p}_2 = 0.23 \tag{9.E.71}$$

b. Test $H_0 : p_1 - p_2 = 0.10$ *vs* $H_a : p_1 - p_2 \neq 0.10$ @ $\alpha = 0.10$

$$n_1 = 220,\ \hat{p}_1 = 0.92 \quad n_2 = 160,\ \hat{p}_2 = 0.78 \tag{9.E.72}$$

11. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.22$ *vs* $H_a : p_1 - p_2 > 0.22$ @ $\alpha = 0.05$

$$n_1 = 90,\ \hat{p}_1 = 0.72 \quad n_2 = 75,\ \hat{p}_2 = 0.40 \tag{9.E.73}$$

b. Test $H_0 : p_1 - p_2 = 0.37$ *vs* $H_a : p_1 - p_2 \neq 0.37$ @ $\alpha = 0.02$

$$n_1 = 425,\ \hat{p}_1 = 0.772\ n_2 = 425,\ \hat{p}_2 = 0.331 \tag{9.E.74}$$

12. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.50$ *vs* $H_a : p_1 - p_2 < 0.50$ @ $\alpha = 0.10$

$$n_1 = 40,\ \hat{p}_1 = 0.65 \quad n_2 = 55,\ \hat{p}_2 = 0.24 \tag{9.E.75}$$

b. Test $H_0 : p_1 - p_2 = 0.30$ *vs* $H_a : p_1 - p_2 \neq 0.30$ @ $\alpha = 0.10$

$$n_1 = 7500,\ \hat{p}_1 = 0.664 \quad n_2 = 1000,\ \hat{p}_2 = 0.319 \tag{9.E.76}$$

Applications

In all the remaining exercsises the samples are sufficiently large (so this need not be checked).

13. Voters in a particular city who identify themselves with one or the other of two political parties were randomly selected and asked if they favor a proposal to allow citizens with proper license to carry a concealed handgun in city parks. The results are:

	Party A	Party B
Sample size, *n*	150	200
Number in favor, *x*	90	140

a. Give a point estimate for the difference in the proportion of all members of **Party A** and all members of **Party B** who favor the proposal.
b. Construct the **95%** confidence interval for the difference, based on these data.
c. Test, at the **5%** level of significance, the hypothesis that the proportion of all members of **Party A** who favor the proposal is less than the proportion of all members of **Party B** who do.
d. Compute the p-value of the test.

14. To investigate a possible relation between gender and handedness, a random sample of 320 adults was taken, with the following results:

	Men	Women
Sample size, n	168	152
Number of left-handed, x	24	9

a. Give a point estimate for the difference in the proportion of all men who are left-handed and the proportion of all women who are left-handed.
b. Construct the 95% confidence interval for the difference, based on these data.
c. Test, at the 5% level of significance, the hypothesis that the proportion of men who are left-handed is greater than the proportion of women who are.
d. Compute the p-value of the test.

15. A local school board member randomly sampled private and public high school teachers in his district to compare the proportions of National Board Certified (NBC) teachers in the faculty. The results were:

	Private Schools	Public Schools
Sample size, n	80	520
Proportion of NBC teachers, $\hat{p}$	0.175	0.150

a. Give a point estimate for the difference in the proportion of all teachers in area public schools and the proportion of all teachers in private schools who are National Board Certified.
b. Construct the 90% confidence interval for the difference, based on these data.
c. Test, at the 10% level of significance, the hypothesis that the proportion of all public school teachers who are National Board Certified is less than the proportion of private school teachers who are.
d. Compute the p-value of the test.

16. In professional basketball games, the fans of the home team always try to distract free throw shooters on the visiting team. To investigate whether this tactic is actually effective, the free throw statistics of a professional basketball player with a high free throw percentage were examined. During the entire last season, this player had 656 free throws, 420 in home games and 236 in away games. The results are summarized below.

	Home	Away
Sample size, n	420	236
Free throw percent, $\hat{p}$	81.5%	78.8%

a. Give a point estimate for the difference in the proportion of free throws made at home and away.
b. Construct the 90% confidence interval for the difference, based on these data.
c. Test, at the 10% level of significance, the hypothesis that there exists a home advantage in free throws.
d. Compute the p-value of the test.

17. Randomly selected middle-aged people in both China and the United States were asked if they believed that adults have an obligation to financially support their aged parents. The results are summarized below.

	China	USA
Sample size, n	1300	150
Number of yes, x	1170	110

Test, at the 1% level of significance, whether the data provide sufficient evidence to conclude that there exists a cultural difference in attitude regarding this question.

18. A manufacturer of walk-behind push mowers receives refurbished small engines from two new suppliers, A and B. It is not uncommon that some of the refurbished engines need to be lightly serviced before they can be fitted into mowers. The mower manufacturer recently received 100 engines from each supplier. In the shipment from A, 13 needed further service. In the shipment from B, 10 needed further service. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that there exists a difference in the proportions of engines from the two suppliers needing service.

Large Data Set Exercises

Large Data Sets are absent

19. Large **Data Sets 6A and 6B** record results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer **Candidate A** for a U.S. Senate seat or prefer some other candidate. Let the population of all voters in **region 1** be denoted **Population 1** and the population of all voters in **region 2** be denoted **Population 2**. Let p_1 be the proportion of voters in **Population 1** who prefer **Candidate A**, and p_2 the proportion in **Population 2** who do.
    a. Find the relevant sample proportions $\hat{p_1}$ and $\hat{p_2}$.
    b. Construct a point estimate for $p_1 - p_2$.
    c. Construct a 95% confidence interval for $p_1 - p_2$.
    d. Test, at the 5% level of significance, the hypothesis that the same proportion of voters in the two regions favor **Candidate A**, against the alternative that a larger proportion in **Population 2** do.
20. Large **Data Set 11** records the results of samples of real estate sales in a certain region in the year 2008 (lines 2 through 536) and in the year 2010 (lines 537 through 1106). Foreclosure sales are identified with a 1 in the second column. Let all real estate sales in the region in 2008 be **Population 1** and all real estate sales in the region in 2010 be **Population 2**.
    a. Use the sample data to construct point estimates $\hat{p_1}$ and $\hat{p_2}$ of the proportions p_1 and p_2 of all real estate sales in this region in 2008 and 2010 that were foreclosure sales. Construct a point estimate of $p_1 - p_2$.
    b. Use the sample data to construct a 90% confidence for $p_1 - p_2$.
    c. Test, at the 10% level of significance, the hypothesis that the proportion of real estate sales in the region in 2010 that were foreclosure sales was greater than the proportion of real estate sales in the region in 2008 that were foreclosure sales. (The default is that the proportions were the same.)

Answers

1. a. $(0.0068, 0.0732)$
   b. $(0.1163, 0.2237)$
2.
3. a. $(0.0210, 0.1030)$
   b. $(0.0001, 0.0319)$
4.
5. a. $Z = 0.996,\ z_{0.10} = 1.282,\ \text{p-value} = 0.1587$, do not reject H_0
   b. $Z = -2.120,\ \pm z_{0.025} = \pm 1.960,\ \text{p-value} = 0.0340$, reject H_0
6.
7. a. $Z = -2.602,\ -z_{0.005} = -2.576,\ \text{p-value} = 0.0047$, reject H_0
   b. $Z = 2.020,\ \pm z_{0.01} = \pm 2.326,\ \text{p-value} = 0.0434$, do not reject H_0
8.
9. a. $Z = -2.85,\ \text{p-value} = 0.0022$, reject H_0
   b. $Z = -2.23,\ \text{p-value} = 0.0258$, do not reject H_0
10.
11. a. $Z = 1.36,\ \text{p-value} = 0.0869$, do not reject H_0
    b. $Z = 2.32,\ \text{p-value} = 0.0204$, do not reject H_0
12.
13. a. -0.10

b. -0.10 ± 0.101
c. $Z = -1.943,\ -z_{0.05} = -1.645$, reject H_0 (fewer in **Party A** favor)
d. **p-value** $= 0.0262$

14.

15. a. 0.025
b. 0.025 ± 0.0745
c. $Z = 0.552,\ z_{0.10} = 1.282$, do not reject H_0 (as many public school teachers are certified)
d. **p-value** $= 0.2912$

16.

17. $Z = 4.498,\ \pm z_{0.005} = \pm 2.576$, reject H_0 (different)

18.

19. a. $\hat{p_1} = 0.355$ and $\hat{p_2} = 0.41$
b. $\hat{p_1} - \hat{p_2} = -0.055$
c. $(-0.1501, 0.0401)$
d. $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 < 0$. Test Statistic: $Z = -1.1335$. Rejection Region: $(-\infty, -1.645]$. Decision: Fail to reject H_0.

9.5 Sample Size Considerations

Basic

1. Estimate the common sample size n of equally sized independent samples needed to estimate $\mu_1 - \mu_2$ as specified when the population standard deviations are as shown.
   a. 90% confidence, to within 3 units, $\sigma_1 = 10$ and $\sigma_2 = 7$
   b. 99% confidence, to within 4 units, $\sigma_1 = 6.8$ and $\sigma_2 = 9.3$
   c. 95% confidence, to within 5 units, $\sigma_1 = 22.6$ and $\sigma_2 = 31.8$
2. Estimate the common sample size n of equally sized independent samples needed to estimate $\mu_1 - \mu_2$ as specified when the population standard deviations are as shown.
   a. 80% confidence, to within 2 units, $\sigma_1 = 14$ and $\sigma_2 = 23$
   b. 90% confidence, to within 0.3 units, $\sigma_1 = 1.3$ and $\sigma_2 = 0.8$
   c. 99% confidence, to within 11 units, $\sigma_1 = 42$ and $\sigma_2 = 37$
3. Estimate the number n of pairs that must be sampled in order to estimate $\mu_d = \mu_1 - \mu_2$ as specified when the standard deviation s_d of the population of differences is as shown.
   a. 80% confidence, to within 6 units, $\sigma_d = 26.5$
   b. 95% confidence, to within 4 units, $\sigma_d = 12$
   c. 90% confidence, to within 5.2 units, $\sigma_d = 11.3$
4. Estimate the number n of pairs that must be sampled in order to estimate $\mu_d = \mu_1 - \mu_2$ as specified when the standard deviation s_d of the population of differences is as shown.
   a. 90% confidence, to within 20 units, $\sigma_d = 75.5$
   b. 95% confidence, to within 11 units, $\sigma_d = 31.4$
   c. 99% confidence, to within 1.8 units, $\sigma_d = 4$
5. Estimate the minimum equal sample sizes $n_1 = n_2$ necessary in order to estimate $p_1 - p_2$ as specified.
   a. 80% confidence, to within 0.05 (five percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.20$ and $p_2 \approx 0.65$
   b. 90% confidence, to within 0.02 (two percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.75$ and $p_2 \approx 0.63$
   c. 95% confidence, to within 0.10 (ten percentage points)

i. when no prior knowledge of p_1 or p_2 is available
ii. when prior studies indicate that $p_1 \approx 0.11$ and $p_2 \approx 0.37$

6. Estimate the minimum equal sample sizes $n_1 = n_2$ necessary in order to estimate $p_1 - p_2$ as specified.
   a. 80% confidence, to within 0.02 (two percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.78$ and $p_2 \approx 0.65$
   b. 90% confidence, to within 0.05 (five percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.12$ and $p_2 \approx 0.24$
   c. 95% confidence, to within 0.10 (ten percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.14$ and $p_2 \approx 0.21$

Applications

7. An educational researcher wishes to estimate the difference in average scores of elementary school children on two versions of a 100-point standardized test, at 99% confidence and to within two points. Estimate the minimum equal sample sizes necessary if it is known that the standard deviation of scores on different versions of such tests is 4.9.
8. A university administrator wishes to estimate the difference in mean grade point averages among all men affiliated with fraternities and all unaffiliated men, with 95% confidence and to within 0.15. It is known from prior studies that the standard deviations of grade point averages in the two groups have common value 0.4. Estimate the minimum equal sample sizes necessary to meet these criteria.
9. An automotive tire manufacturer wishes to estimate the difference in mean wear of tires manufactured with an experimental material and ordinary production tire, with 90% confidence and to within 0.5 mm. To eliminate extraneous factors arising from different driving conditions the tires will be tested in pairs on the same vehicles. It is known from prior studies that the standard deviations of the differences of wear of tires constructed with the two kinds of materials is 1.75 mm. Estimate the minimum number of pairs in the sample necessary to meet these criteria.
10. To assess to the relative happiness of men and women in their marriages, a marriage counselor plans to administer a test measuring happiness in marriage to n randomly selected married couples, record the their test scores, find the differences, and then draw inferences on the possible difference. Let μ_1 and μ_2 be the true average levels of happiness in marriage for men and women respectively as measured by this test. Suppose it is desired to find a 90% confidence interval for estimating $\mu_d = \mu_1 - \mu_2$ to within two test points. Suppose further that, from prior studies, it is known that the standard deviation of the differences in test scores is $\sigma_d \approx 10$. What is the minimum number of married couples that must be included in this study?
11. A journalist plans to interview an equal number of members of two political parties to compare the proportions in each party who favor a proposal to allow citizens with a proper license to carry a concealed handgun in public parks. Let p_1 and p_2 be the true proportions of members of the two parties who are in favor of the proposal. Suppose it is desired to find a 95% confidence interval for estimating $p_1 - p_2$ to within 0.05. Estimate the minimum equal number of members of each party that must be sampled to meet these criteria.
12. A member of the state board of education wants to compare the proportions of National Board Certified (NBC) teachers in private high schools and in public high schools in the state. His study plan calls for an equal number of private school teachers and public school teachers to be included in the study. Let p_1 and p_2 be these proportions. Suppose it is desired to find a 99% confidence interval that estimates $p_1 - p_2$ to within 0.05.
    a. Supposing that both proportions are known, from a prior study, to be approximately 0.15, compute the minimum common sample size needed.
    b. Compute the minimum common sample size needed on the supposition that nothing is known about the values of p_1 and p_2.

Answers

1. a. $n_1 = n_2 = 45$
   b. $n_1 = n_2 = 56$
   c. $n_1 = n_2 = 234$

2.
3. a. $n_1 = n_2 = 33$
   b. $n_1 = n_2 = 35$
   c. $n_1 = n_2 = 13$
4.
5. a. i. $n_1 = n_2 = 329$
      ii. $n_1 = n_2 = 255$
   b. i. $n_1 = n_2 = 3383$
      ii. $n_1 = n_2 = 2846$
   c. i. $n_1 = n_2 = 193$
      ii. $n_1 = n_2 = 128$
6.
7. $n_1 = n_2 \approx 80$
8.
9. $n_1 = n_2 \approx 34$
10.
11. $n_1 = n_2 \approx 769$

This page titled 9.E: Two-Sample Problems (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **9.E: Two-Sample Problems (Exercises)** has no license indicated.

CHAPTER OVERVIEW

10: Linear Regression and Correlation

Regression analysis is a statistical process for estimating the relationships among variables and includes many techniques for modeling and analyzing several variables. When the focus is on the relationship between a dependent variable and one or more independent variables.

10.1: Introduction to Linear Regression and Correlation

10.1.1: Linear Equations

10.1.1E: Linear Equations (Exercises)

10.1.2: Scatter Plots

10.1.2E: Scatter Plots (Exercises)

10.2: The Regression Equation and Correlation Coefficient

10.2E: The Regression Equation (Exercise)

10.3: Testing for Significance Linear Correlation

10.3E: Testing the Significance of the Correlation Coefficient (Exercises)

10.4: Prediction

10.4E: Prediction (Exercises)

10.E: Linear Regression and Correlation (Exercises)

This page titled 10: Linear Regression and Correlation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1: Introduction to Linear Regression and Correlation

CHAPTER OBJECTIVES

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

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

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, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.



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)

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables. In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

This page titled 10.1: Introduction to Linear Regression and Correlation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1.1: 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*. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 10.1.1.1

The following examples are linear equations.

$$y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

Exercise 10.1.1.1

Is the following an example of a linear equation?

$$y = -0.125 - 3.5x$$

Answer

yes

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 10.1.1.2

Graph the equation $y = -1 + 2x$.

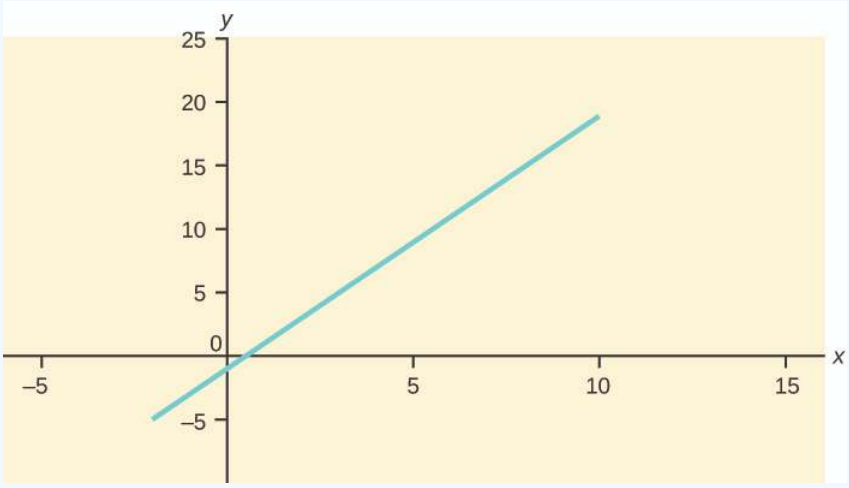


Figure 10.1.1.1.

Exercise 10.1.1.2

Is the following an example of a linear equation? Why or why not?

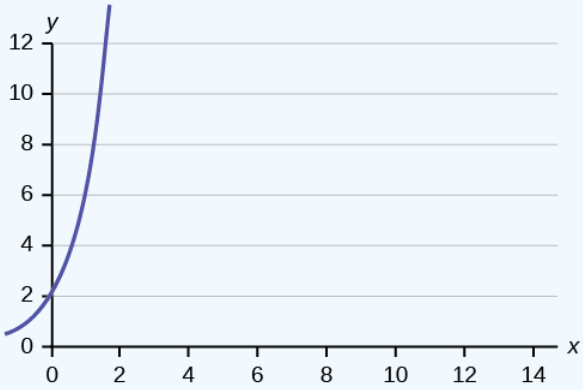


Figure 10.1.1.2.

Answer

No, the graph is not a straight line; therefore, it is not a linear equation.

Example 10.1.1.3

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: $y = 31.50 + 32x$

Exercise 10.1.1.3

Emma’s Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Answer

$y = 50 + 20x$

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.

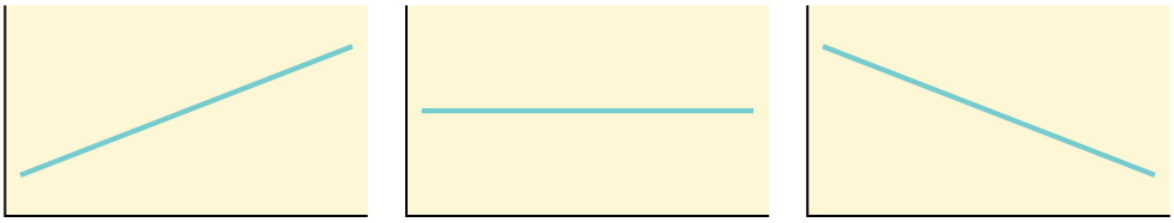


Figure 10.1.1.3:. Three possible graphs of $y = a + bx$ (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 10.1.1.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the 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 Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Exercise 10.1.1.4

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

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 Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is 25 ($a = 25$). At the start of a visit, Ethan charges a one-time fee of $25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Ethan earns $20 for each hour he works.

Summary

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the x variable (b is called a coefficient) is called the **slope**. The constant a is called the y-intercept.

The **slope of a line** is a value that describes the rate of change between the 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.

Formula Review

$y = a + bx$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

This page titled 10.1.1: Linear Equations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1.1E: Linear Equations (Exercises)

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

Exercise 12.2.5

What are the dependent and independent variables?

Answer

dependent variable: fee amount; independent variable: time

Exercise 12.2.6

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Exercise 12.2.7

Graph the equation from Exercise.

Answer

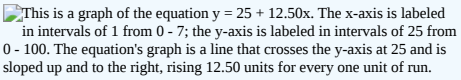


Figure **10.1.1*E*.4**.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

Exercise 12.2.8

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Exercise 12.2.9

Graph the equation from Exercise.

Answer

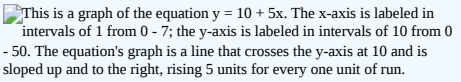


Figure **10.1.1*E*.5**.

Exercise 12.2.10

Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Exercise 12.2.11

Which of the following equations are linear?

a. $y = 6x + 8$
b. $y + 7 = 3x$
c. $y - x = 8x^2$
d. $4y = 8$

Answer

$y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

Exercise 12.2.12

Does the graph show a linear equation? Why or why not?

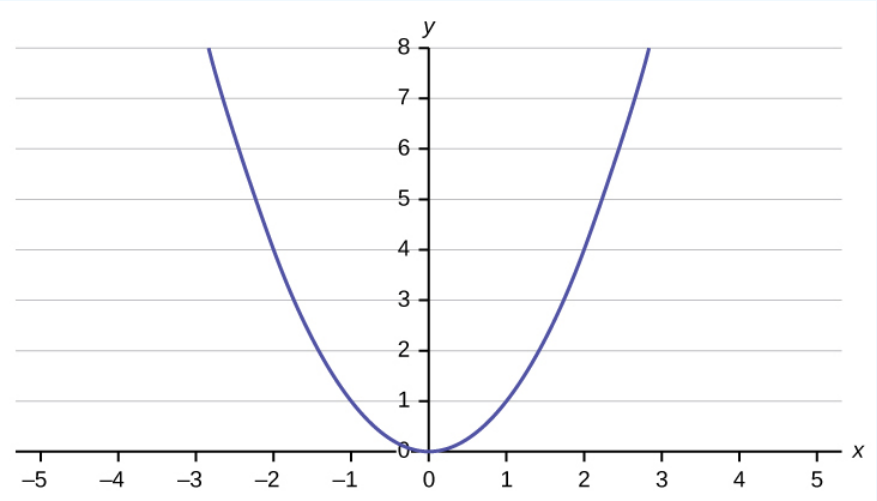


Figure 10.1.1E.6.

Table contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

https://stats.libretexts.org/@go/page/48598

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Exercise 12.2.13

Use the columns "year" and "# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Answer

The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

Exercise 12.2.14

What are the independent and dependent variables?

Exercise 12.2.15

What is the *y*-intercept and what is the slope? Interpret them using complete sentences.

Answer

The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of \$50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges \$100 for each hour they clean.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12,000x$.

Exercise 12.2.16

What are the independent and dependent variables?

Exercise 12.2.17

How many pounds of soil does the shoreline lose in a year?

Answer

12,000 pounds of soil

Exercise 12.2.18

What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15–1.5x$ where x is the number of hours passed in an eight-hour day of trading.

Exercise 12.2.19

What are the slope and *y*-intercept? Interpret their meaning.

Answer

The slope is -1.5 ($b = -1.5$). This means the stock is losing value at a rate of \$1.50 per hour. The y-intercept is \$15 ($a = 15$). This means the price of stock before the trading day was \$15.

Exercise 12.2.19

If you owned this stock, would you want a positive or negative slope? Why?

This page titled 10.1.1E: Linear Equations (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

10.1.2: Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables *x* and *y*. The most common and easiest way is a *scatter plot*. The following example illustrates a scatter plot.

Example 10.1.2.1

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let $x =$ the year and let $y =$ the number of m-commerce users, in millions.

Table **10.1.2.1**: Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

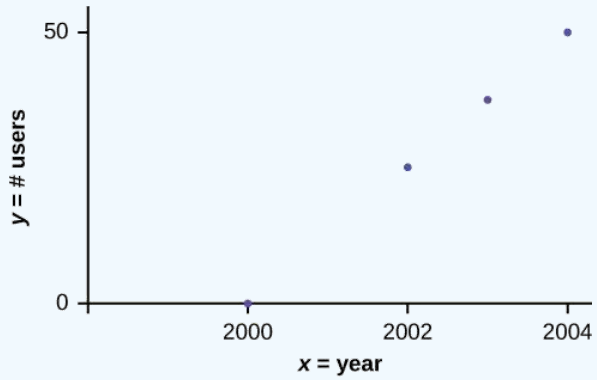


Figure **10.1.2.1**: Scatter plot showing the number of m-commerce users (in millions) by year.

To create a scatter plot

a. Enter your X data into list L1 and your Y data into list L2.
b. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
c. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
d. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
e. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
f. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
g. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Exercise 10.1.2.1

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Answer

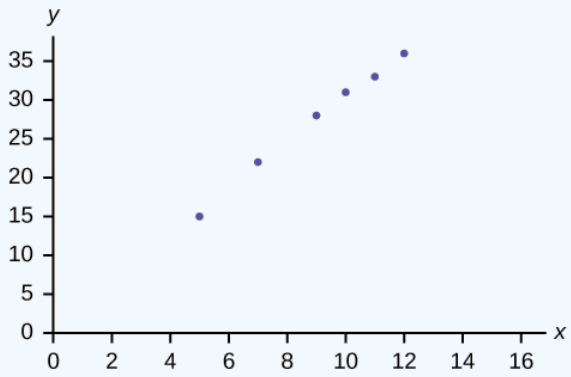


***Figure* 10.1.2.2**

Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatter plot shows the *direction of a relationship* between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the *strength of the relationship* by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatter plot, you want to notice the *overall pattern* and any *deviations* from the pattern. The following scatterplot examples illustrate these concepts.

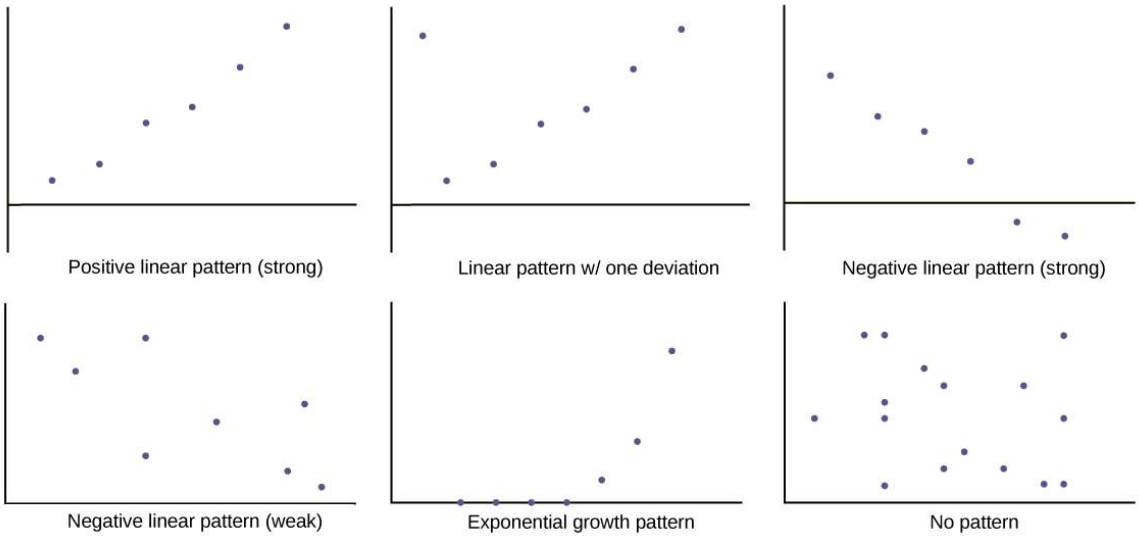


Figure **10.1.2.3**:

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

Summary

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

This page titled 10.1.2: Scatter Plots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1.2E: Scatter Plots (Exercises)

Exercise 10.1.2E.1

Does the scatter plot appear linear? Strong or weak? Positive or negative?

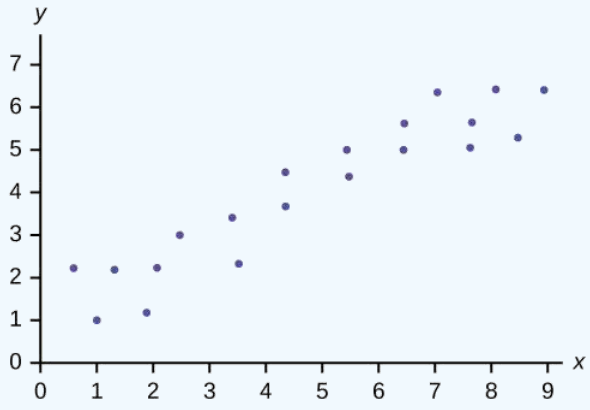


Figure 10.1.2E.4

Answer

The data appear to be linear with a strong, positive correlation.

Exercise 10.1.2E.3

Does the scatter plot appear linear? Strong or weak? Positive or negative?

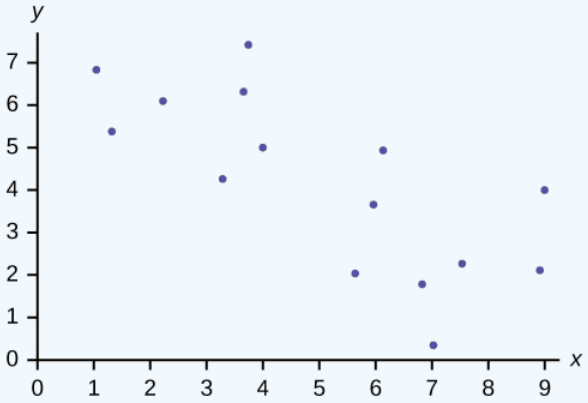


Figure 10.1.2E.5

Exercise 10.1.2E.4

Does the scatter plot appear linear? Strong or weak? Positive or negative?

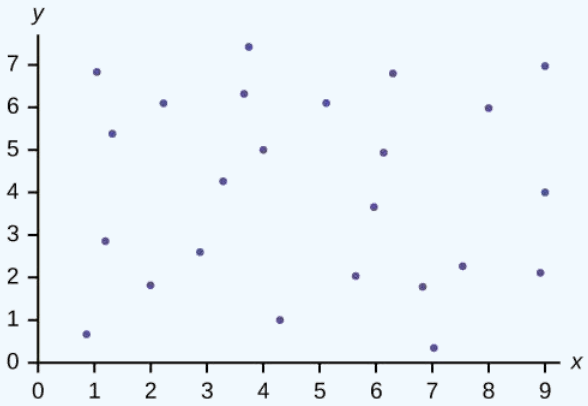


Figure **10.1.2*E*.6**

Answer

The data appear to have no correlation.

This page titled 10.1.2E: Scatter Plots (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.2: The Regression Equation and Correlation Coefficient

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a Line of Best Fit **or** Least-Squares Line.

COLLABORATIVE EXERCISE

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y -intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Example 10.2.1

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 random student if you know the third exam score?

1a: Table showing the scores on the final exam based on scores from the third exam.

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

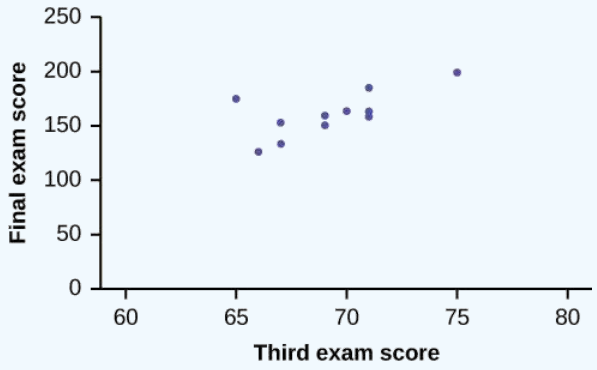


Figure **10.2.1**: Scatter plot showing the scores on the final exam based on scores from the third exam.

Exercise 10.2.1

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in Table show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

X (depth in feet)	Y (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Answer

$\hat{y} = 127.24–1.11x$

At 110 feet, a diver could dive for only five minutes.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a least-squares regression line to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read "y **hat**" and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.

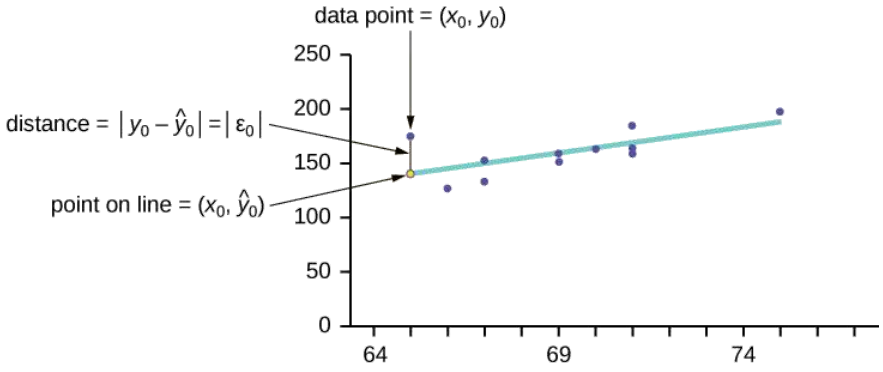


Figure **10.2.2**

The term $y_0–\hat{y}_0 = \varepsilon_0$ is called the **"error" or** residual. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram in Figure, $y_0–\hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

$\varepsilon =$ the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + ... + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2 \quad (10.2.1)$$

Equation 10.2.1 is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. 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:

$$\hat{y} = a + bx \quad (10.2.2)$$

where

- $a = \bar{y} - b\bar{x}$ and
- $b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$.

The sample means of the x values and the x values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r\left(\frac{s_y}{s_x}\right)$ where $s_y =$ the standard deviation of the y values and $s_x =$ the standard deviation of the x values. r is the correlation coefficient, which is discussed in the next section.

Least Square Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line** .

> **Note**
>
> Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third-exam/final-exam example is as follows:

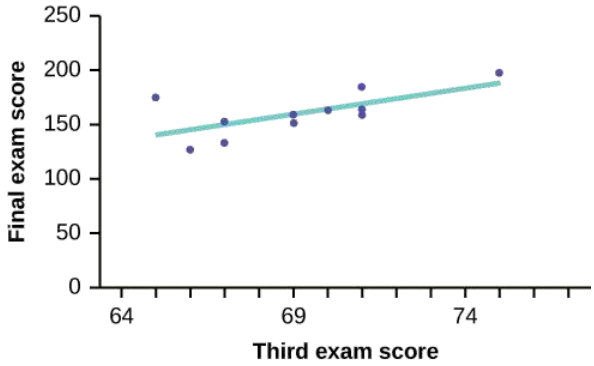


Figure **10.2.3**

The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

$$\hat{y} = -173.51 + 4.83x \tag{10.2.3}$$

REMINDER

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.** You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

Understanding Slope

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

Slope: The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Using the Linear Regression T Test: LinRegTTest

a. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x, y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
b. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.)
c. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
d. On the next line, at the prompt β or ρ, highlight "$\neq 0$" and press ENTER
e. Leave the line for "RegEq:" blank
f. Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen

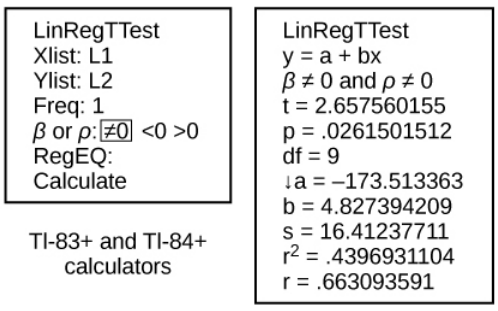


Figure 10.2.4

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says $y = a + bx$. Scroll down to find the values $a = -173.513$, and $b = 4.8273$; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$

The two items at the bottom are $r_2 = 0.43969$ and $r = 0.663$. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1
3. On the input screen for PLOT 1, highlight **On**, and press ENTER
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
5. Indicate Xlist: L1 and Ylist: L2
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
8. To graph the best-fit line, press the "$Y =$" key and type the equation $-173.5 + 4.83X$ into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

Note

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

a. Make sure you have done the scatter plot. Check it on your screen.
b. Go to LinRegTTest and enter the lists.
c. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
d. Press $Y =$ (you will see the regression equation).
e. Press GRAPH. The line will be drawn."

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y. The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{10.2.4}$$

where $n =$ the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

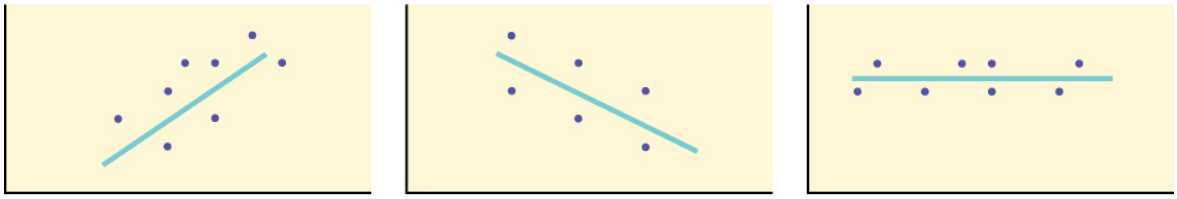
What the VALUE of r tells us:

- 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 x and y. Values of r close to –1 or to +1 indicate a stronger linear relationship between x and y.
- If $r = 0$ there is absolutely no linear relationship between 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)**.
- The sign of r is the same as the sign of the slope, b, of the best-fit line.

> **Note**
>
> Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."**



(a) Positive correlation (b) Negative correlation (c) Zero correlation

Figure 10.2.5: (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called *the coefficient of determination* and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation $(1 - 0.44 = 0.56)$ in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Summary

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of y and the estimated value of y. The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2, is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

Glossary

Coefficient of Correlation

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{10.2.5}$$

where n is the number of data points. The coefficient cannot be more than 1 or less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y.

This page titled 10.2: The Regression Equation and Correlation Coefficient is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.2E: The Regression Equation (Exercise)

Use the following information to answer the next five exercises. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Exercise 12.4.2

Draw a scatter plot of the data.

Exercise 12.4.3

Use regression to find the equation for the line of best fit.

Answer

$\hat{y} = 2.23 + 1.99x$

Exercise 12.4.4

Draw the line of best fit on the scatter plot.

Exercise 12.4.5

What is the slope of the line of best fit? What does it represent?

Answer

The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another $1.99 million in pay each year.

Exercise 12.4.6

What is the y-intercept of the line of best fit? What does it represent?

Exercise 12.4.7

What does an r value of zero mean?

Answer

It means that there is no correlation between the data sets.

Exercise 12.4.8

When $n = 2$ and $r = 1$, are the data significant? Explain.

Exercise 12.4.9

When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

This page titled 10.2E: The Regression Equation (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.3: Testing for Significance Linear Correlation

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together. We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
- $\rho =$ population correlation coefficient (unknown)
- $r =$ sample correlation coefficient (known; calculated from sample data)

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 x and y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

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".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

NOTE

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

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. The two methods are equivalent and give the same result.

- **Method 1: Using the p-value**

- **Method 1: Using the p-value**
- **Method 2: Using a table of critical values**

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

NOTE

Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

Using the TI83, 83+, 84, 84+ CALCULATOR

To calculate the p-value using LinRegTTEST:

On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "$\neq 0$"

The output screen shows the p-value on the line that reads "$p =$".

(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level ($\alpha = 0.05$):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p-value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Calculation Notes:

- You will use technology to calculate the p-value. The following describes the calculations to compute the test statistics and the p-value:
- The p-value is calculated using a t-distribution with $n - 2$ degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r.
- The p-value is the combined area in both tails.

An alternative way to calculate the p-value (p) given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: p-value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software).
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$.

- Decision: Reject the Null Hypothesis H_0
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example 10.3.1

Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and $+0.632$. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

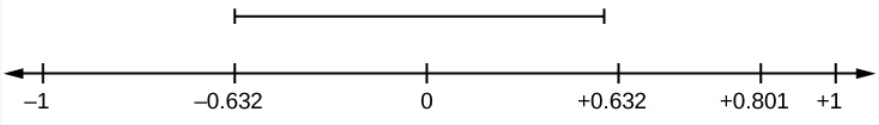


Figure 10.3.1. r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.

Exercise 10.3.1

For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points and the critical value is 0.576. Can the line be used for prediction? Why or why not?

Answer

If the scatter plot looks linear then, yes, the line can be used for prediction, because $r >$ the positive critical value.

Example 10.3.2

Suppose you computed $r = –0.624$ with 14 data points. $df = 14–2 = 12$. The critical values are -0.532 and 0.532. Since $-0.624 < -0.532$, r is significant and the line can be used for prediction

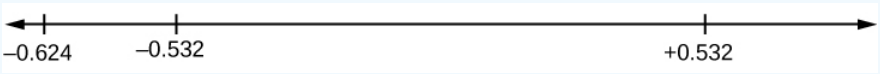


Figure 10.3.2. $r = -0.624 - 0.532$. Therefore, r is significant.

Exercise 10.3.2

For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical value is 0.666. Can the line be used for prediction? Why or why not?

Answer

No, the line cannot be used for prediction, because $r <$ the positive critical value.

Example 10.3.3

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant, and the line should not be used for prediction.



Figure 10.3.3. $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

? Exercise 10.3.3

For a given line of best fit, you compute that $r = -0.7204$ using $n = 8$ data points, and the critical value is $= 0.707$. Can the line be used for prediction? Why or why not?

Answer

Yes, the line can be used for prediction, because $r <$ the negative critical value.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: critical value method

Consider the third exam/final exam example. The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points. Can the regression line be used for prediction? **Given a third-exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$
- $\alpha = 0.05$
- Use the "95% Critical Value" table for r with $df = n - 2 = 11 - 2 = 9$.
- The critical values are -0.602 and $+0.602$
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion:There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

✓ Example 10.3.4

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

a. $r = -0.567$ and the sample size, n, is 19. The $df = n - 2 = 17$. The critical value is -0.456. $-0.567 < -0.456$ so r is significant.
b. $r = 0.708$ and the sample size, n, is 9. The $df = n - 2 = 7$. The critical value is 0.666. $0.708 > 0.666$ so r is significant.
c. $r = 0.134$ and the sample size, n, is 14. The $df = 14 - 2 = 12$. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so r is not significant.
d. $r = 0$ and the sample size, n, is five. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

? Exercise 10.3.4

For a given line of best fit, you compute that $r = 0$ using $n = 100$ data points. Can the line be used for prediction? Why or why not?

Answer

No, the line cannot be used for prediction no matter what the sample size is.

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatter plot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.

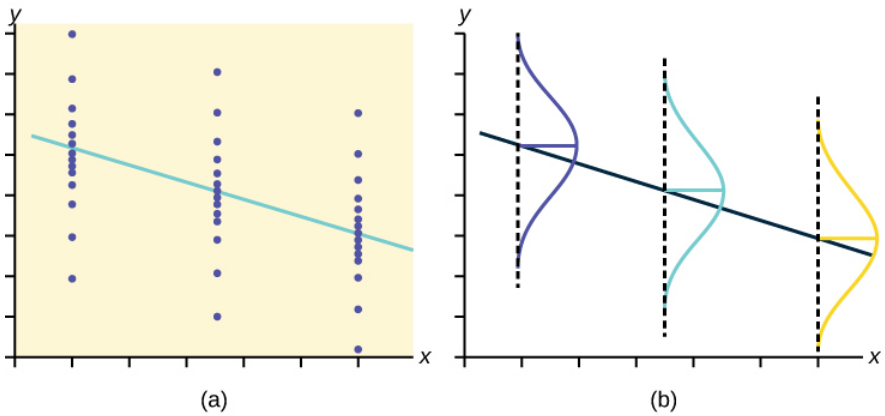


Figure **10.3.4**. The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

Summary

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear** In the population, there is a linear relationship that models the average value of y for different values of x.
- **Independent** The residuals are assumed to be independent.
- **Normal** The y values are distributed normally for any value of x.
- **Equal variance** The standard deviation of the y values is equal for each x value.
- **Random** The data are produced from a well-designed random sample or randomized experiment.

The slope b and intercept a of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y, σ, use the standard deviation of the residuals, s. $s = \sqrt{\frac{SEE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis $H_0 : \rho =$ *hypothesized value*, use a linear regression t-test. The

most common null hypothesis is $H_0 : \rho = 0$ which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit:

$$\hat{y} = a + bx \tag{10.3.1}$$

where

$$a = y\text{-intercept} \tag{10.3.2}$$

$$b = \text{slope} \tag{10.3.3}$$

Standard deviation of the residuals:

$$s = \sqrt{\frac{SSE}{n-2}} \tag{10.3.4}$$

where

$$SSE = \text{sum of squared errors} \tag{10.3.5}$$

$$n = \text{the number of data points} \tag{10.3.6}$$

This page titled 10.3: Testing for Significance Linear Correlation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.3E: Testing the Significance of the Correlation Coefficient (Exercises)

Exercise 10.3E.5

When testing the significance of the correlation coefficient, what is the null hypothesis?

Exercise 10.3E.6

When testing the significance of the correlation coefficient, what is the alternative hypothesis?

Answer

$H_a : \rho \neq 0$

Exercise 10.3E.7

If the level of significance is 0.05 and the p-value is 0.04, what conclusion can you draw?

This page titled 10.3E: Testing the Significance of the Correlation Coefficient (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.4: Prediction

Recall the third exam/final exam example. We examined the scatter plot and showed that the correlation coefficient is significant. 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.

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-values 65 and 75, 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.

Example 10.4.1

Recall the third exam/final exam example.

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 not be reliable.)

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.

The process of predicting inside of the observed x values observed in the data is called *interpolation*. The process of predicting outside of the observed x-values observed in the data is called *extrapolation*.

Exercise 10.4.1

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$$\hat{y} = 72.5 + 2.8x$$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Answer

86.5

Summary

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least squares regression line to make predictions about your data.

References

1. Data from the Centers for Disease Control and Prevention.
2. Data from the National Center for HIV, STD, and TB Prevention.
3. Data from the United States Census Bureau. Available online at www.census.gov/compendia/stat...atalities.html
4. Data from the National Center for Health Statistics.

This page titled 10.4: Prediction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.4E: Prediction (Exercises)

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as follows:

$$\hat{y} = 101.32 + 2.48x \tag{10.4E.1}$$

where $\hat{y}$ is in thousands of dollars.

Exercise 12.6.2

What would you predict the sales to be on day 60?

Answer

$250,120

Exercise 12.6.3

What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is as follows:

$$\hat{y} = 1350 - 1.2x \tag{10.4E.2}$$

where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

Exercise 12.6.4

How many acres will be left to mow after 20 hours of work?

Answer

1,326 acres

Exercise 12.6.5

How many acres will be left to mow after 100 hours of work?

Exercise 12.6.7

How many hours will it take to mow all of the lawns? (When is $\hat{y} = 0$?)

Answer

1,125 hours, or when $x = 1,125$

Table contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

? Exercise 12.6.8

Graph “year” versus “# AIDS cases diagnosed” (plot the scatter plot). Do not include pre-1981 data.

? Exercise 12.6.9

Perform linear regression. What is the linear equation? Round to the nearest whole number.

Answer

Check student’s solution.

Exercise 12.6.10

Write the equations:

a. Linear equation: _____
b. $a =$ _____
c. $b =$ _____
d. $r =$ _____
e. $n =$ _____

Exercise 12.6.11

Solve.

a. When $x = 1985$, $\hat{y} =$ _____
b. When $x = 1990$, $\hat{y} =$_____
c. When $x = 1970$, $\hat{y} =$_____ Why doesn't this answer make sense?

Answer

a. When $x = 1985$, $\hat{y} = 25,52$
b. When $x = 1990$, $\hat{y} = 34,275$
c. When $x = 1970$, $\hat{y} = -725$ Why doesn't this answer make sense? The range of x values was 1981 to 2002; the year 1970 is not in this range. The regression equation does not apply, because predicting for the year 1970 is extrapolation, which requires a different process. Also, a negative number does not make sense in this context, where we are predicting AIDS cases diagnosed.

Exercise 12.6.11

Does the line seem to fit the data? Why or why not?

Exercise 12.6.12

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Answer

Also, the correlation $r = 0.4526$. If r is compared to the value in the 95% Critical Values of the Sample Correlation Coefficient Table, because $r > 0.423$, r is significant, and you would think that the line could be used for prediction. But the scatter plot indicates otherwise.

Exercise 12.6.13

Plot the two given points on the following graph. Then, connect the two points to form the regression line.

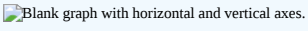


Figure 10.4*E*.1.

Obtain the graph on your calculator or computer.

Exercise 12.6.14

Write the equation: $\hat{y} =$ _____

Answer

$\hat{y} = 3,448,225 + 1750x$

Exercise 12.6.15

Hand draw a smooth curve on the graph that shows the flow of the data.

Exercise 12.6.16

Does the line seem to fit the data? Why or why not?

Answer

There was an increase in AIDS cases diagnosed until 1993. From 1993 through 2002, the number of AIDS cases diagnosed declined each year. It is not appropriate to use a linear regression line to fit to the data.

Exercise 12.6.17

Do you think a linear fit is best? Why or why not?

Exercise 12.6.18

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Answer

Since there is no linear association between year and # of AIDS cases diagnosed, it is not appropriate to calculate a linear correlation coefficient. When there is a linear association and it is appropriate to calculate a correlation, we cannot say that one variable "causes" the other variable.

Exercise 12.6.19

Graph "year" vs. "# AIDS cases diagnosed." Do not include pre-1981. Label both axes with words. Scale both axes.

Exercise 12.6.20

Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Write the linear equation, rounding to four decimal places:

Answer

We don't know if the pre-1981 data was collected from a single year. So we don't have an accurate x value for this figure.

Regression equation: Misplaced '#'

	Coefficients
Intercept	–3,448,225
X Variable 1	1,749.777

Exercise 12.6.21

Calculate the following:

a. $a =$ _____
b. $b =$ _____
c. correlation = _____

d. $n =$ _____

This page titled 10.4E: Prediction (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.E: Linear Regression and Correlation (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

12.1: Introduction

12.2: Linear Equations

Q 12.2.1

For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

S 12.2.1

a. independent variable: age; dependent variable: fatalities
b. independent variable: # of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

Q 12.2.2

Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000 with an additional $125 added per percentage point from 81–99%	$6,500 with an additional $125 added per percentage point from 101–119%	$9,500 with an additional $125 added per percentage point starting at 121%

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

12.3: Scatter Plots

Q 12.3.1

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. Table shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

S 12.3.1

Check student's solution.

Q 12.3.2

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Q 12.3.3

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

S 12.3.3

For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

Q 12.3.4

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

Q 12.3.5

If there are 15 data points in a set of data, what is the number of degree of freedom?

S 12.3.5

13

12.4: The Regression Equation

Q 12.4.1

What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

Q 12.4.2

Explain what it means when a correlation has an r^2 of 0.72.

S 12.4.2

It means that 72% of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

Q 12.4.3

Can a coefficient of determination be negative? Why or why not?

12.5: Testing the Significance of the Correlation Coefficient

Q 12.5.1

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

S 12.5.1

We do not reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.

Q 12.5.2

If there are 15 data points in a set of data, what is the number of degree of freedom?

12.6: Prediction

Q 12.6.1

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

a. For each age group, pick the midpoint of the interval for the *x* value. (For the 75+ group, use 80.)

b. Using “ages” as the independent variable and “Number of driver deaths per 100,000” as the dependent variable, make a scatter plot of the data.
c. Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Predict the number of deaths for ages 40 and 60.
f. Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
g. What is the slope of the least squares (best-fit) line? Interpret the slope.

S 12.6.1

a.

Age	Number of Driver Deaths per 100,000
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

b. Check student’s solution.
c. $\hat{y} = 35.5818045 - 0.19182491x$
d. $r = -0.57874$

For four df and alpha $= 0.05$, the LinRegTTest gives $p\text{-value} = 0.2288$ so we do not reject the null hypothesis; there is not a significant linear relationship between deaths and age.

Using the table of critical values for the correlation coefficient, with four *df*, the critical value is 0.811. The correlation coefficient $r = -0.57874$ is not less than –0.811, so we do not reject the null hypothesis.

e. if age = 40, $\hat{y}\ (\text{deaths}) = 35.5818045–0.19182491(40) = 27.9$

if age = 60, $\hat{y}\ (\text{deaths}) = 35.5818045–0.19182491(60) = 24.1$

f. For entire dataset, there is a linear relationship for the ages up to age 74. The oldest age group shows an increase in deaths from the prior group, which is not consistent with the younger ages.
g. $\text{slope} = -0.19182491$

Q 12.6.2

Table shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to part e the same as the values in Table that correspond to those years?
g. Use the two points in part e to plot the least squares line on your graph from part b.
h. Based on the data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the data?
j. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Q 12.6.3

The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition ten, for various pages is given in Table

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.

c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page ten and on page 70.
f. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
g. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
h. Is the least squares line valid for page 200? Why or why not?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

S 12.6.3

a. We wonder if the better discounts appear earlier in the book so we select page as X and discount as Y.
b. Check student's solution.
c. $\hat{y} = 17.21757 - 0.01412x$
d. $r = -0.2752$

For seven df and $\alpha = 0.05$, using LinRegTTest $p\text{-value} = 0.4736$ so we do not reject; there is a not a significant linear relationship between page and discount.

Using the table of critical values for the correlation coefficient, with seven df, the critical value is 0.666. The correlation coefficient $xi = -0.2752$ is not less than 0.666 so we do not reject.

e. page 10: 17.08 page 70: 16.23
f. There is not a significant linear correlation so it appears there is no relationship between the page and the amount of the discount.
g. page 200: 14.39
h. No, using the regression equation to predict for page 200 is extrapolation.
i. $\text{slope} = -0.01412$

As the page number increases by one page, the discount decreases by $0.01412

Q 12.6.4

Table gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

a. Decide which variable should be the independent variable and which should be the dependent variable.

b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from part f different from the chart values?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

Q 12.6.5

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Use the least-squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

S 12.6.5

a. Year is the independent or x variable; the number of letters is the dependent or y variable.
b. Check student's solution.
c. no
d. $\hat{y} = 47.03 - 0.0216x$
e. -0.4280
f. 6; 5
g. No, the relationship does not appear to be linear; the correlation is not significant.

h. current year: 2013: 3.55 or four letters; this is not an appropriate use of the least squares line. It is extrapolation.

12.7: Outliers

Q 12.7.1

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

a. Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Based on the data in Table, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

Q 12.7.2

Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.

Percent return: 74; 66; 81; 52; 73; 62; 52; 45; 62; 46; 60; 46; 38

Percent new: 5; 6; 8; 11; 12; 15; 16; 17; 18; 18; 19; 20; 20

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator’s regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?

S 12.7.2

a. Check student's solution.
b. Check student's solution.
c. The slope of the regression line is -0.3179 with a y-intercept of 32.966. In context, the y-intercept indicates that when there are no returning sparrow hawks, there will be almost 31% new sparrow hawks, which doesn't make sense since if there are no returning birds, then the new percentage would have to be 100% (this is an example of why we do not extrapolate). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 0.3179%.
d. If we examine $r2$, we see that only 50.238% of the variation in the percent of new birds is explained by the model and the correlation coefficient, $r = 0.71$ only indicates a somewhat strong correlation between returning and new percentages.
e. The ordered pair $(66, 6)$ generates the largest residual of 6.0. This means that when the observed return percentage is 66%, our observed new percentage, 6%, is almost 6% less than the predicted new value of 11.98%. If we remove this data pair, we see only an adjusted slope of -0.2723 and an adjusted intercept of 30.606. In other words, even though this data generates the largest residual, it is not an outlier, nor is the data pair an influential point.
f. If there are 70% returning birds, we would expect to see $y = -0.2723(70) + 30.606 = 0.115$ or 11.5% new birds in the colony.

Q 12.7.3

The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries.

Yearly wine consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. Do the data provide convincing evidence that there is a linear relationship between the amount of alcohol consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

Q 12.7.4

The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

S 12.7.4

a. Check student's solution.
b. Check student's solution.
c. We have a slope of -1.4946 with a y-intercept of 193.88. The slope, in context, indicates that for each additional minute added to the swim time, the heart rate will decrease by 1.5 beats per minute. If the student is not swimming at all, the y-intercept indicates that his heart rate will be 193.88 beats per minute. While the slope has meaning (the longer it takes to swim 2,000 meters, the less effort the heart puts out), the y-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.
d. Since only 1.5% of the heart rate variation is explained by this regression equation, we must conclude that this association is not explained with a linear relationship.
e. The point $(34.72, 124)$ generates the largest residual of -11.82. This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. When this point is removed, the slope becomes 1.6914 with the y-intercept changing to 83.694. While the linear association is still very weak, we see that the removed data pair can be considered an influential point in the sense that the y-intercept becomes more meaningful.

Q 12.7.5

A researcher is investigating whether non-white minorities commit a disproportionate number of homicides. He uses demographic data from Detroit, MI to compare homicide rates and the number of the population that are white males.

White Males	Homicide rate per 100,000 people
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

a. Use your calculator to construct a scatter plot of the data. What should the independent variable be? Why?
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
c. Discuss what the following mean in context.
   i. The slope of the regression equation
   ii. The y-intercept of the regression equation
   iii. The correlation r
   iv. The coefficient of determination $r2$.
d. Do the data provide convincing evidence that there is a linear relationship between the number of white males in the population and the homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

Q 12.7.6

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Using the data to determine the linear-regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.

S 12.7.6

If we remove the two service academies (the tuition is $0.00), we construct a new regression equation of $y = -0.0009x + 160$ with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the data set.

12.8: Regression (Distance from School)

12.9: Regression (Textbook Cost)

12.10: Regression (Fuel Efficiency)

This page titled 10.E: Linear Regression and Correlation (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **12.E: Linear Regression and Correlation (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

11: Chi-Square Tests

11.1: Chi-Square Tests for Independence

11.2: Chi-Square One-Sample Goodness-of-Fit Tests

11: Chi-Square Tests is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

11.1: Chi-Square Tests for Independence

Learning Objectives

- To understand what chi-square distributions are.
- To understand how to use a chi-square test to judge whether two factors are independent.

Chi-Square Distributions

As you know, there is a whole family of t-distributions, each one specified by a parameter called the degrees of freedom, denoted df. Similarly, all the chi-square distributions form a family, and each of its members is also specified by a parameter df, the number of degrees of freedom. Chi is a Greek letter denoted by the symbol χ and chi-square is often denoted by χ^2.

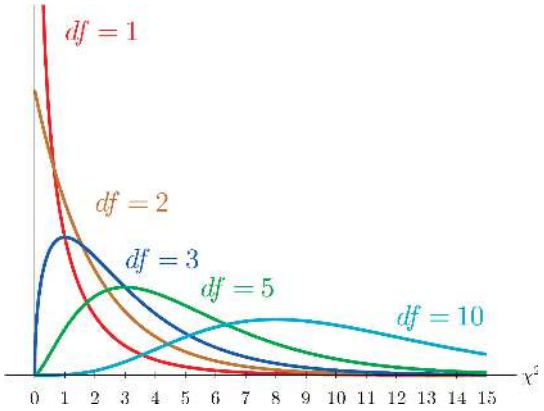


Figure 11.1.1: Many χ Distributions

Figure 11.1.1 shows several χ-square distributions for different degrees of freedom. A chi-square random variable is a random variable that assumes only positive values and follows a χ-square distribution.

Definition: critical value

The value of the chi-square random variable χ^2 with $df = k$ that cuts off a right tail of area c is denoted χ^2_c and is called a critical value (Figure 11.1.2).

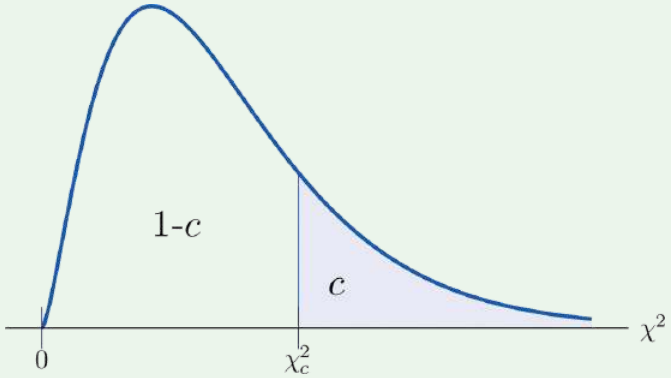
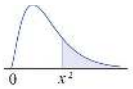


Figure 11.1.2: χ^2_c Illustrated

Figure 11.1.3 below gives values of χ^2_c for various values of c and under several chi-square distributions with various degrees of freedom.



Critical Values of Chi-Square Distributions

	x^2 Right-Tail Area									
df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.96	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 11.1.3: Critical Values of Chi-Square Distributions

Tests for Independence

Hypotheses tests encountered earlier in the book had to do with how the numerical values of two population parameters compared. In this subsection we will investigate hypotheses that have to do with whether or not two random variables take their values independently, or whether the value of one has a relation to the value of the other. Thus the hypotheses will be expressed in words, not mathematical symbols. We build the discussion around the following example.

There is a theory that the gender of a baby in the womb is related to the baby's heart rate: baby girls tend to have higher heart rates. Suppose we wish to test this theory. We examine the heart rate records of 40 babies taken during their mothers' last prenatal checkups before delivery, and to each of these 40 randomly selected records we compute the values of two random measures: 1) gender and 2) heart rate. In this context these two random measures are often called factors. Since the burden of proof is that heart rate and gender are related, not that they are unrelated, the problem of testing the theory on baby gender and heart rate can be formulated as a test of the following hypotheses:

H_0 : Baby gender and baby heart rate are independent

vs.

H_a : Baby gender and baby heart rate are not independent

The factor gender has two natural categories or levels: boy and girl. We divide the second factor, heart rate, into two levels, low and high, by choosing some heart rate, say 145 beats per minute, as the cutoff between them. A heart rate below 145 beats per minute will be considered low and 145 and above considered high. The 40 records give rise to a 2×2 *contingency table*. By adjoining row totals, column totals, and a grand total we obtain the table shown as Table 11.1.1. The four entries in boldface type are counts of observations from the sample of $n = 40$. There were 11 girls with low heart rate, 17 boys with low heart rate, and so on. They form the *core* of the expanded table.

Table 11.1.1: Baby Gender and Heart Rate

		Heart Rate		
		Low	High	Row Total
Gender	Girl	11	7	18
	Boy	17	5	22
Column Total		28	12	Total = 40

In analogy with the fact that the probability of independent events is the product of the probabilities of each event, if heart rate and gender were independent then we would expect the number in each core cell to be close to the product of the row total R and column total C of the row and column containing it, divided by the sample size n. Denoting such an expected number of observations E, these four expected values are:

- 1st row and 1st column: $E = (R \times C)/n = 18 \times 28/40 = 12.6$
- 1st row and 2nd column: $E = (R \times C)/n = 18 \times 12/40 = 5.4$
- 2nd row and 1st column: $E = (R \times C)/n = 22 \times 28/40 = 15.4$
- 2nd row and 2nd column: $E = (R \times C)/n = 22 \times 12/40 = 6.6$

We update Table 11.1.1 by placing each expected value in its corresponding core cell, right under the observed value in the cell. This gives the updated table Table 11.1.2.

Table 11.1.2: Updated Baby Gender and Heart Rate

		Heart Rate		
		Low	High	Row Total
Gender	Girl	$O = 11$ $E = 12.6$	$O = 7$ $E = 5.4$	$R = 18$
	Boy	$O = 17$ $E = 15.4$	$O = 5$ $E = 6.6$	$R = 22$
Column Total		$C = 28$	$C = 12$	$n = 40$

A measure of how much the data deviate from what we would expect to see if the factors really were independent is the sum of the squares of the difference of the numbers in each core cell, or, standardizing by dividing each square by the expected number in the cell, the sum $\sum (O - E)^2/E$. We would reject the null hypothesis that the factors are independent only if this number is large, so the test is right-tailed. In this example the random variable $\sum (O - E)^2/E$ has the chi-square distribution with one degree of freedom. If we had decided at the outset to test at the 10% level of significance, the critical value defining the rejection region would be, reading from Figure 11.1.3, $\chi^2_\alpha = \chi^2_{0.10} = 2.706$, so that the rejection region would be the interval $[2.706, \infty)$. When we compute the value of the standardized test statistic we obtain

$$\sum \frac{(O-E)^2}{E} = \frac{(11-12.6)^2}{12.6} + \frac{(7-5.4)^2}{5.4} + \frac{(17-15.4)^2}{15.4} + \frac{(5-6.6)^2}{6.6} = 1.231$$

Since $1.231 < 2.706$, the decision is not to reject H_0. See Figure 11.1.4. The data do not provide sufficient evidence, at the 10% level of significance, to conclude that heart rate and gender are related.

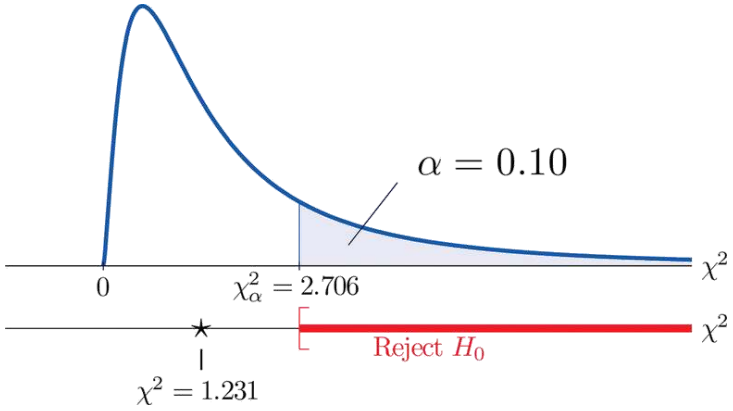


***Figure* 11.1.4**: *Baby Gender Prediction*

With this specific example in mind, now turn to the general situation. In the general setting of testing the independence of two factors, call them Factor 1 and Factor 2, the hypotheses to be tested are

$$H_0 : \text{The two factors are independent}$$
$$vs.$$
$$H_a : \text{The two factors are not independent}$$

As in the example each factor is divided into a number of categories or levels. These could arise naturally, as in the boy-girl division of gender, or somewhat arbitrarily, as in the high-low division of heart rate. Suppose Factor 1 has I levels and Factor 2 has J levels. Then the information from a random sample gives rise to a general $I \times J$ contingency table, which with row totals, column totals, and a grand total would appear as shown in Table 11.1.3. Each cell may be labeled by a pair of indices (i, j). O_{ij} stands for the observed count of observations in the cell in row i and column j, R_i for the ith row total and C_j for the jth column total. To simplify the notation we will drop the indices so Table 11.1.3 becomes Table 11.1.4. Nevertheless it is important to keep in mind that the Os, the Rs and the Cs, though denoted by the same symbols, are in fact different numbers.

Table 11.1.3: General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O_{11}	$\cdots$	O_{1j}	$\cdots$	O_{1J}	R_1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O_{i1}	$\cdots$	O_{ij}	$\cdots$	O_{iJ}	R_i
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O_{I1}	$\cdots$	O_{Ij}	$\cdots$	O_{IJ}	R_I
Column Total		C_1	$\cdots$	C_j	$\cdots$	C_J	n

Table 11.1.4: Simplified General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O	$\cdots$	O	$\cdots$	O	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O	$\cdots$	O	$\cdots$	O	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O	$\cdots$	O	$\cdots$	O	R
Column Total		C	$\cdots$	C	$\cdots$	C	n

As in the example, for each core cell in the table we compute what would be the expected number E of observations if the two factors were independent. E is computed for each core cell (each cell with an O in it) of Table 11.1.4 by the rule applied in the example:

$$E = R \times Cn$$

where R is the row total and C is the column total corresponding to the cell, and n is the sample size

After the expected number is computed for every cell, Table 11.1.4 is updated to form Table 11.1.5 by inserting the computed value of E into each core cell.

Table 11.1.5: Updated General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O E	$\cdots$	O E	$\cdots$	O E	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O E	$\cdots$	O E	$\cdots$	O E	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O E	$\cdots$	O E	$\cdots$	O E	R
Column Total		C	$\cdots$	C	$\cdots$	C	n

Here is the test statistic for the general hypothesis based on Table 11.1.5, together with the conditions that it follow a chi-square distribution.

Test Statistic for Testing the Independence of Two Factors

$$\chi^2 = \sum (O - E)^2 E$$

where the sum is over all core cells of the table.

If

1. the two study factors are independent, and
2. the observed count O of each cell in Table 11.1.5 is at least 5,

then χ^2 approximately follows a chi-square distribution with $df = (I-1) \times (J-1)$ degrees of freedom.

The same five-step procedures, either the critical value approach or the p-value approach, that were introduced in Section 8.1 and Section 8.3 are used to perform the test, which is always right-tailed.

Example 11.1.1

A researcher wishes to investigate whether students' scores on a college entrance examination (CEE) have any indicative power for future college performance as measured by GPA. In other words, he wishes to investigate whether the factors CEE and GPA are independent or not. He randomly selects $n = 100$ students in a college and notes each student's score on the entrance examination and his grade point average at the end of the sophomore year. He divides entrance exam scores into two levels and grade point averages into three levels. Sorting the data according to these divisions, he forms the contingency table shown as Table 11.1.6, in which the row and column totals have already been computed.

Table 11.1.6: *CEE versus GPA Contingency Table*

		GPA			
		< 2.7	2.7 to 3.2	> 3.2	Row Total
CEE	< 1800	35	12	5	52
	≥ 1800	6	24	18	48
Column Total		41	36	23	Total = 100

Test, at the 1% level of significance, whether these data provide sufficient evidence to conclude that CEE scores indicate future performance levels of incoming college freshmen as measured by GPA.

Solution

We perform the test using the critical value approach, following the usual five-step method outlined at the end of Section 8.1.

- **Step 1**. The hypotheses are

$$H_0 : \text{CEE and GPA are independent factors}$$
$$vs.$$
$$H_a : \text{CEE and GPA are not independent factors}$$

- **Step 2**. The distribution is chi-square.
- **Step 3**. To compute the value of the test statistic we must first computed the expected number for each of the six core cells (the ones whose entries are boldface):
  - 1st row and 1st column: $E = (R \times C)/n = 41 \times 52/100 = 21.32$
  - 1st row and 2nd column: $E = (R \times C)/n = 36 \times 52/100 = 18.72$
  - 1st row and 3rd column: $E = (R \times C)/n = 23 \times 52/100 = 11.96$
  - 2nd row and 1st column: $E = (R \times C)/n = 41 \times 48/100 = 19.68$
  - 2nd row and 2nd column: $E = (R \times C)/n = 36 \times 48/100 = 17.28$
  - 2nd row and 3rd column: $E = (R \times C)/n = 23 \times 48/100 = 11.04$

Table 11.1.6 is updated to Table 11.1.6.

Table 11.1.7: Updated CEE versus GPA Contingency Table

		GPA			
		< 2.7	2.7 to 3.2	> 3.2	Row Total
CEE	< 1800	$O = 35$ $E = 21.32$	$O = 12$ $E = 18.72$	$O = 5$ $E = 11.96$	$R = 52$
	≥ 1800	$O = 6$ $E = 19.68$	$O = 24$ $E = 17.28$	$O = 18$ $E = 11.04$	$R = 48$
Column Total		$C = 41$	$C = 36$	$C = 23$	$n = 100$

The test statistic is

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(35-21.32)^2}{21.32} + \frac{(12-18.72)^2}{18.72} + \frac{(5-11.96)^2}{11.96} + \frac{(6-19.68)^2}{19.68} \\ &\quad + \frac{(24-17.28)^2}{17.28} + \frac{(18-11.04)^2}{11.04} \\ &= 31.75\end{aligned}$$

- **Step 4**. Since the CEE factor has two levels and the GPA factor has three, $I = 2$ and $J = 3$. Thus the test statistic follows the chi-square distribution with $df = (2-1) \times (3-1) = 2$ degrees of freedom.

  Since the test is right-tailed, the critical value is $\chi^2_{0.01}$. Reading from Figure 7.1.6 "Critical Values of Chi-Square Distributions", $\chi^2_{0.01} = 9.210$, so the rejection region is $[9.210, \infty)$.

- **Step 5**. Since $31.75 > 9.21$ the decision is to reject the null hypothesis. See Figure 11.1.5. The data provide sufficient evidence, at the 1% level of significance, to conclude that CEE score and GPA are not independent: the entrance exam score has predictive power.

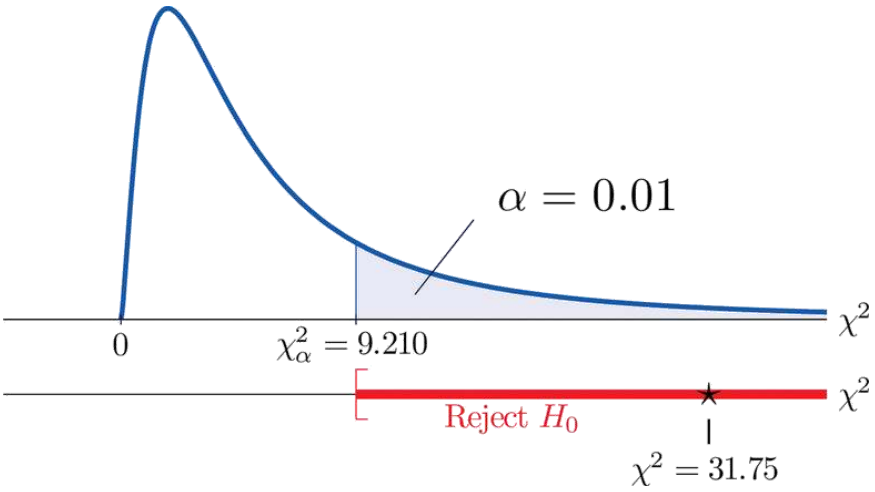


Figure 11.1.5: "Example 11.1.1"

Key Takeaway

- Critical values of a chi-square distribution with degrees of freedom df are found in Figure 7.1.6.
- A chi-square test can be used to evaluate the hypothesis that two random variables or factors are independent.

This page titled 11.1: Chi-Square Tests for Independence is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **11.1: Chi-Square Tests for Independence** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

11.2: Chi-Square One-Sample Goodness-of-Fit Tests

Learning Objectives

- To understand how to use a chi-square test to judge whether a sample fits a particular population well.

Suppose we wish to determine if an ordinary-looking six-sided die is fair, or balanced, meaning that every face has probability $1/6$ of landing on top when the die is tossed. We could toss the die dozens, maybe hundreds, of times and compare the actual number of times each face landed on top to the expected number, which would be $1/6$ of the total number of tosses. We wouldn't expect each number to be exactly $1/6$ of the total, but it should be close. To be specific, suppose the die is tossed $n = 60$ times with the results summarized in Table 11.2.1. For ease of reference we add a column of expected frequencies, which in this simple example is simply a column of $10s$. The result is shown as Table 11.2.2. In analogy with the previous section we call this an "updated" table. A measure of how much the data deviate from what we would expect to see if the die really were fair is the sum of the squares of the differences between the observed frequency O and the expected frequency E in each row, or, standardizing by dividing each square by the expected number, the sum

$$\frac{\Sigma(O-E)^2}{E}$$

If we formulate the investigation as a test of hypotheses, the test is

$$H_0 : \text{The die is fair}$$
$$vs.$$
$$H_a : \text{The die is not fair}$$

Table 11.2.1: Die Contingency Table

Die Value	Assumed Distribution	Observed Frequency
1	1/6	9
2	1/6	15
3	1/6	9
4	1/6	8
5	1/6	6
6	1/6	13

Table 11.2.2: Updated Die Contingency Table

Die Value	Assumed Distribution	Observed Freq.	Expected Freq.
1	1/6	9	10
2	1/6	15	10
3	1/6	9	10
4	1/6	8	10
5	1/6	6	10
6	1/6	13	10

We would reject the null hypothesis that the die is fair only if the number $\frac{\Sigma(O-E)^2}{E}$ is large, so the test is right-tailed. In this example the random variable $\frac{\Sigma(O-E)^2}{E}$ has the chi-square distribution with five degrees of freedom. If we had decided at the

outset to test at the 10% level of significance, the critical value defining the rejection region would be, reading from Figure 7.1.6, $\chi^2_\alpha = \chi^2_{0.10} = 9.236$, so that the rejection region would be the interval

$$[9.236, \infty)$$

. When we compute the value of the standardized test statistic using the numbers in the last two columns of Table 11.2.2, we obtain

$$\begin{aligned}\sum \frac{(O-E)^2}{E} &= \frac{(-1)^2}{10} + \frac{(5)^2}{10} + \frac{(-1)^2}{10} + \frac{(-2)^2}{10} + \frac{(-4)^2}{10} + \frac{(3)^2}{10} \\ &= 0.1 + 2.5 + 0.1 + 0.4 + 1.6 + 0.9 \\ &= 5.6\end{aligned}$$

Since $5.6 < 9.236$ the decision is not to reject H_0. See Figure 11.2.1. The data do not provide sufficient evidence, at the 10% level of significance, to conclude that the die is loaded.

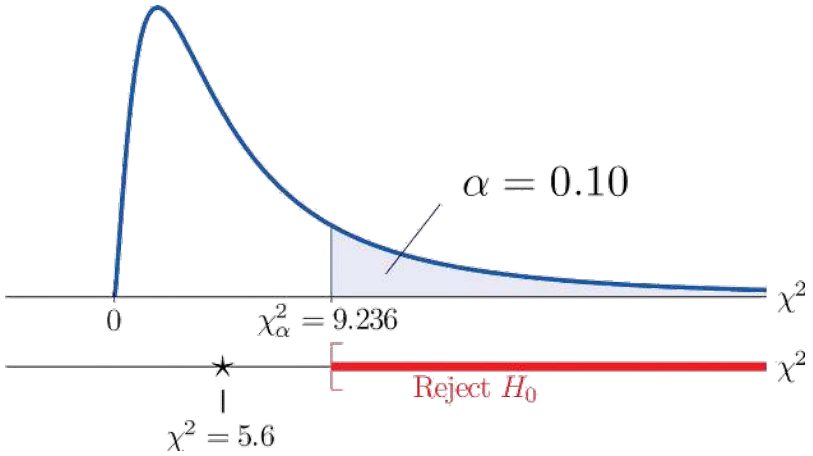


Figure 11.2.1: Balanced Die

In the general situation we consider a discrete random variable that can take I different values, $x_1, x_2, \cdots, x_I$, for which the default assumption is that the probability distribution is

x	x_1	x_2	$\cdots$	x_I
$P(x)$	p_1	p_2	$\cdots$	p_I

We wish to test the hypotheses:

$$H_0 : \text{The assumed probability distribution for X is valid}$$
$$vs.$$
$$H_a : \text{The assumed probability distribution for X is not valid}$$

We take a sample of size n and obtain a list of observed frequencies. This is shown in Table 11.2.3. Based on the assumed probability distribution we also have a list of assumed frequencies, each of which is defined and computed by the formula

$$Ei = n \times pi$$

Table 11.2.3: General Contingency Table

Factor Levels	Assumed Distribution	Observed Frequency
1	p_1	O_1
2	p_2	O_2
⋮	⋮	⋮
I	p_I	O_I

Table 11.2.3 is updated to Table 11.2.4 by adding the expected frequency for each value of X. To simplify the notation we drop indices for the observed and expected frequencies and represent Table 11.2.4 by Table 11.2.5.

Table 11.2.4: Updated General Contingency Table

Factor Levels	Assumed Distribution	Observed Freq.	Expected Freq.
1	p_1	O_1	E_1
2	p_2	O_2	E_2
⋮	⋮	⋮	⋮
I	p_I	O_I	E_I

Table 11.2.5: Simplified Updated General Contingency Table

Factor Levels	Assumed Distribution	Observed Freq.	Expected Freq.
1	p_1	O	E
2	p_2	O	E
⋮	⋮	⋮	⋮
I	p_I	O	E

Here is the test statistic for the general hypothesis based on Table 11.2.5, together with the conditions that it follow a chi-square distribution.

Test Statistic for Testing Goodness of Fit to a Discrete Probability Distribution

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where the sum is over all the rows of the table (one for each value of X).

If

1. the true probability distribution of X is as assumed, and
2. the observed count O of each cell in Table 11.2.5 is at least 5,

then χ^2 approximately follows a chi-square distribution with $df = I - 1$ degrees of freedom.

The test is known as a goodness-of-fit χ^2 test since it tests the null hypothesis that the sample fits the assumed probability distribution well. It is always right-tailed, since deviation from the assumed probability distribution corresponds to large values of χ^2.

Testing is done using either of the usual five-step procedures.

Example 11.2.1

Table 11.2.6 shows the distribution of various ethnic groups in the population of a particular state based on a decennial U.S. census. Five years later a random sample of 2,500 residents of the state was taken, with the results given in Table 11.2.7 (along with the probability distribution from the census year). Test, at the 1% level of significance, whether there is sufficient evidence in the sample to conclude that the distribution of ethnic groups in this state five years after the census had changed from that in the census year.

Table 11.2.6: Ethnic Groups in the Census Year

Ethnicity	White	Black	Amer.-Indian	Hispanic	Asian	Others
Proportion	0.743	0.216	0.012	0.012	0.008	0.009

Table 11.2.7: Sample Data Five Years After the Census Year

Ethnicity	Assumed Distribution	Observed Frequency
White	0.743	1732
Black	0.216	538
American-Indian	0.012	32
Hispanic	0.012	42
Asian	0.008	133
Others	0.009	23

Solution

We test using the critical value approach.

- **Step 1**. The hypotheses of interest in this case can be expressed as

$$H_0 : \text{The distribution of ethnic groups has not changed}$$
$$vs.$$
$$H_a : \text{The distribution of ethnic groups has changed}$$

- **Step 2**. The distribution is chi-square.
- **Step 3**. To compute the value of the test statistic we must first compute the expected number for each row of Table 11.2.7. Since $n = 2500$, using the formula $E_i = n \times p_i$ and the values of p_i from either Table 11.2.6 or Table 11.2.7,

$$E_1 = 2500 \times 0.743 = 1857.5$$
$$E_2 = 2500 \times 0.216 = 540$$
$$E_3 = 2500 \times 0.012 = 30$$
$$E_4 = 2500 \times 0.012 = 30$$
$$E_5 = 2500 \times 0.008 = 20$$
$$E_6 = 2500 \times 0.009 = 22.5$$

Table 11.2.7 is updated to Table 11.2.8.

Table 11.2.8: Observed and Expected Frequencies Five Years After the Census Year

Ethnicity	Assumed Dist.	Observed Freq.	Expected Freq.
White	0.743	1732	1857.5
Black	0.216	538	540
American-Indian	0.012	32	30
Hispanic	0.012	42	30
Asian	0.008	133	20
Others	0.009	23	22.5

The value of the test statistic is

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(1732-1857.5)^2}{1857.5} + \frac{(538-540)^2}{540} + \frac{(32-30)^2}{30} \\ &\quad + \frac{(42-30)^2}{30} + \frac{(133-20)^2}{20} + \frac{(23-22.5)^2}{22.5} \\ &= 651.881\end{aligned}$$

Since the random variable takes six values, $I = 6$. Thus the test statistic follows the chi-square distribution with $df = 6 - 1 = 5$ degrees of freedom.

Since the test is right-tailed, the critical value is $\chi^2_{0.01}$. Reading from Figure 7.1.6, $\chi^2_{0.01} = 15.086$, so the rejection region is $[15.086, \infty)$.

Since $651.881 > 15.086$ the decision is to reject the null hypothesis. See Figure 11.2.2. The data provide sufficient evidence, at the 1% level of significance, to conclude that the ethnic distribution in this state has changed in the five years since the U.S. census.

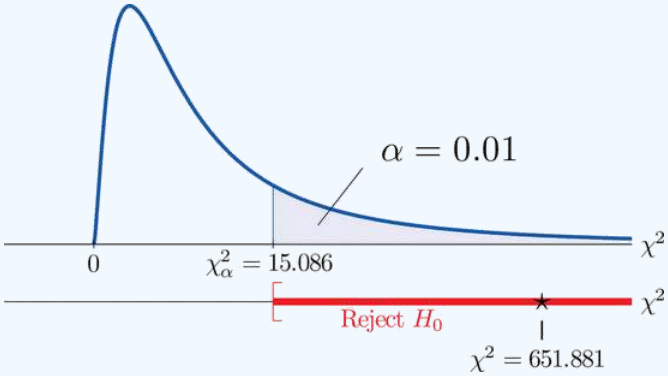


Figure 11.2.2: "Example 11.2.1"

Key Takeaway

- The chi-square goodness-of-fit test can be used to evaluate the hypothesis that a sample is taken from a population with an assumed specific probability distribution.

This page titled 11.2: Chi-Square One-Sample Goodness-of-Fit Tests is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **11.2: Chi-Square One-Sample Goodness-of-Fit Tests** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

CHAPTER OVERVIEW

12: Analysis of Variance

12.1: F-Tests
12.2: F-Tests in One-Way ANOVA

12: Analysis of Variance is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

12.1: F-Tests

Learning Objectives

- To understand what F-distributions are.
- To understand how to use an F-test to judge whether two population variances are equal.

F-Distributions

Another important and useful family of distributions in statistics is the family of F-distributions. Each member of the F-distribution family is specified by a pair of parameters called degrees of freedom and denoted df_1 and df_2. Figure **12.1.1** shows several F-distributions for different pairs of degrees of freedom. An F random variable is a random variable that assumes only positive values and follows an F-distribution.

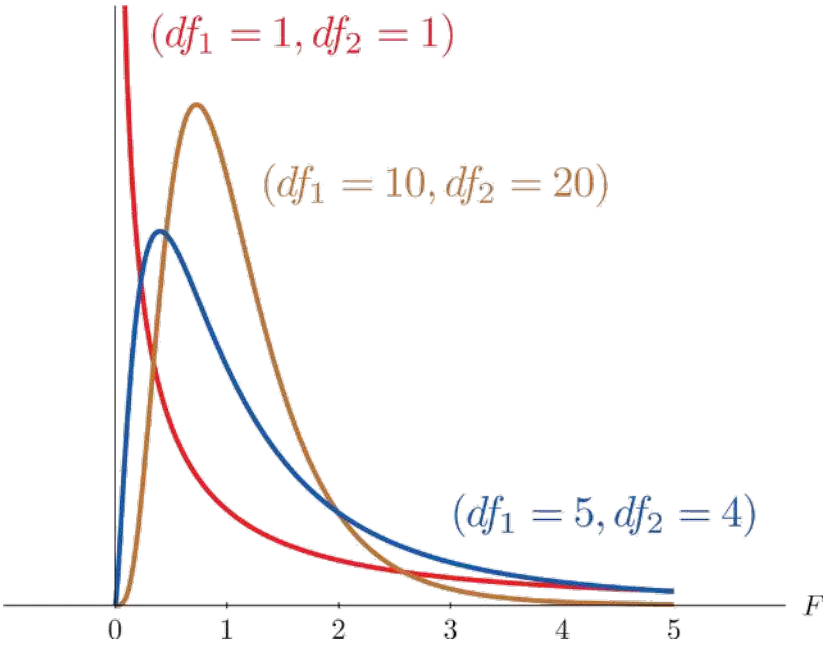


Figure **12.1.1**: Many *F-Distributions*

The parameter df_1 is often referred to as the numerator degrees of freedom and the parameter df_2 as the denominator degrees of freedom. It is important to keep in mind that they are not interchangeable. For example, the F-distribution with degrees of freedom $df_1 = 3$ and $df_2 = 8$ is a different distribution from the F-distribution with degrees of freedom $df_1 = 8$ and $df_2 = 3$.

Definition: critical value

The value of the F random variable F with degrees of freedom df_1 and df_2 that cuts off a right tail of area c is denoted F_c and is called a critical value (Figure **12.1.2**).

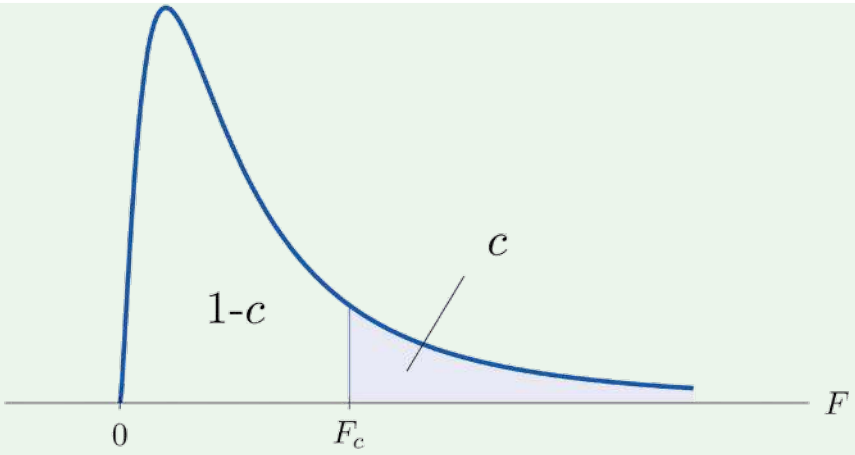


Figure 12.1.2: F_c Illustrated

Tables containing the values of F_c are given in Chapter 11. Each of the tables is for a fixed collection of values of c, either 0.900, 0.950, 0.975, 0.990, and 0.995 (yielding what are called "lower" critical values), or 0.005, 0.010, 0.025, 0.050, and 0.100 (yielding what are called "upper" critical values). In each table critical values are given for various pairs (df_1, df_2). We illustrate the use of the tables with several examples.

Example 12.1.1: an F random variable

Suppose F is an F random variable with degrees of freedom $df_1 = 5$ and $df_2 = 4$. Use the tables to find

1. $F_{0.10}$
2. $F_{0.95}$

Solution

1. The column headings of all the tables contain $df_1 = 5$. Look for the table for which 0.10 is one of the entries on the extreme left (a table of upper critical values) and that has a row heading $df_2 = 4$ in the left margin of the table. A portion of the relevant table is provided. The entry in the intersection of the column with heading $df_1 = 5$ and the row with the headings 0.10 and $df_2 = 4$, which is shaded in the table provided, is the answer,

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯	5	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.005	4	⋯	⋯	⋯	22.5	⋯
0.01	4	⋯	⋯	⋯	15.5	⋯
0.025	4	⋯	⋯	⋯	9.36	⋯
0.05	4	⋯	⋯	⋯	6.26	⋯
0.10	4	⋯	⋯	⋯	4.05	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮

2. Look for the table for which 0.95 is one of the entries on the extreme left (a table of lower critical values) and that has a row heading $df_2 = 4$ in the left margin of the table. A portion of the relevant table is provided. The entry in the intersection of the column with heading $df_1 = 5$ and the row with the headings 0.95 and $df_2 = 4$, which is shaded in the table provided, is the answer,

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯	5	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.90	4	⋯	⋯	⋯	0.28	⋯
0.95	4	⋯	⋯	⋯	0.19	⋯
0.975	4	⋯	⋯	⋯	0.14	⋯
0.99	4	⋯	⋯	⋯	0.09	⋯
0.995	4	⋯	⋯	⋯	0.06	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Example 12.1.2

Suppose F is an F random variable with degrees of freedom $df_1 = 2$ and $df_2 = 20$. Let $\alpha = 0.05$. Use the tables to find

1. F_{α}
2. $F_{\alpha/2}$
3. $F_{1-\alpha}$
4. $F_{1-\alpha/2}$

Solution

1. The column headings of all the tables contain $df_1 = 2$. Look for the table for which $\alpha = 0.05$ is one of the entries on the extreme left (a table of upper critical values) and that has a row heading $df_2 = 20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1 = 2$ and the row with the headings 0.05 and $df_2 = 20$ is the answer,

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.005	20	⋯	6.99	⋯
0.01	20	⋯	5.85	⋯
0.025	20	⋯	4.46	⋯
0.05	20	⋯	3.49	⋯
0.10	20	⋯	2.59	⋯
⋮	⋮	⋮	⋮	⋮

2. Look for the table for which $\alpha/2 = 0.025$ is one of the entries on the extreme left (a table of upper critical values) and that has a row heading $df_2 = 20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1 = 2$ and the row with the headings 0.025 and $df_2 = 20$ is the answer, $F_{0.025} = 4.46$.

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.005	20	⋯	6.99	⋯
0.01	20	⋯	5.85	⋯
0.025	20	⋯	4.46	⋯
0.05	20	⋯	3.49	⋯
0.10	20	⋯	2.59	⋯
⋮	⋮	⋮	⋮	⋮

3. Look for the table for which $1-\alpha=0.95$ is one of the entries on the extreme left (a table of lower critical values) and that has a row heading $df_2=20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1=2$ and the row with the headings 0.95 and $df_2=20$ is the answer, $F_{0.95}=0.05$.

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.90	20	⋯	0.11	⋯
0.95	20	⋯	0.05	⋯
0.975	20	⋯	0.03	⋯
0.99	20	⋯	0.01	⋯
0.995	20	⋯	0.01	⋯
⋮	⋮	⋮	⋮	⋮

4. Look for the table for which $1-\alpha/2=0.975$ is one of the entries on the extreme left (a table of lower critical values) and that has a row heading $df_2=20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1=2$ and the row with the headings 0.975 and $df_2=20$ is the answer, $F_{0.975}=0.03$.

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.90	20	⋯	0.11	⋯
0.95	20	⋯	0.05	⋯
0.975	20	⋯	0.03	⋯
0.99	20	⋯	0.01	⋯
0.995	20	⋯	0.01	⋯
⋮	⋮	⋮	⋮	⋮

A fact that sometimes allows us to find a critical value from a table that we could not read otherwise is:

If $F_u(r,s)$ denotes the value of the F-distribution with degrees of freedom $df_1=r$ and $df_2=s$ that cuts off a right tail of area u, then

$$F_c(k,l)=\frac{1}{F_{1-c}(l,k)}$$

Example 12.1.3

Use the tables to find

1. $F_{0.01}$ for an F random variable with $df_1 = 13$ and $df_2 = 8$.
2. $F_{0.975}$ for an F random variable with $df_1 = 40$ and $df_2 = 10$.

Solution

1. There is no table with $df_1 = 13$, but there is one with $df_1 = 8$. Thus we use the fact that

$$F_{0.01}(13, 8) = \frac{1}{F_{0.99}(8, 13)}$$

   Using the relevant table we find that $F_{0.99}(8, 13) = 0.18$, hence $F_{0.01}(13, 8) = 0.18^{-1} = 5.556$.
2. There is no table with $df_1 = 40$, but there is one with $df_1 = 10$. Thus we use the fact that

$$F_{0.975}(40, 10) = \frac{1}{F_{0.025}(10, 40)}$$

   Using the relevant table we find that $F_{0.025}(10, 40) = 3.31$, hence $F_{0.975}(40, 10) = 3.31^{-1} = 0.302$.

F-Tests for Equality of Two Variances

In Chapter 9 we saw how to test hypotheses about the difference between two population means μ_1 and μ_2. In some practical situations the difference between the population standard deviations σ_1 and σ_2 is also of interest. Standard deviation measures the variability of a random variable. For example, if the random variable measures the size of a machined part in a manufacturing process, the size of standard deviation is one indicator of product quality. A smaller standard deviation among items produced in the manufacturing process is desirable since it indicates consistency in product quality.

For theoretical reasons it is easier to compare the squares of the population standard deviations, the population variances σ_1^2 and σ_2^2. This is not a problem, since $\sigma_1 = \sigma_2$ precisely when $\sigma_1^2 = \sigma_2^2$, $\sigma_1 < \sigma_2$ precisely when $\sigma_1^2 < \sigma_2^2$, and $\sigma_1 > \sigma_2$ precisely when $\sigma_1^2 > \sigma_2^2$.

The null hypothesis always has the form $H_0 : \sigma_1^2 = \sigma_2^2$. The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : \sigma_1^2 > \sigma_2^2$	Right-tailed
$H_a : \sigma_1^2 < \sigma_2^2$	Left-tailed
$H_a : \sigma_1^2 \neq \sigma_2^2$	Two-tailed

Just as when we test hypotheses concerning two population means, we take a random sample from each population, of sizes n_1 and n_2, and compute the sample standard deviations s_1 and s_2. In this context the samples are always independent. The populations themselves must be normally distributed.

Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Variances

$$F = \frac{s_1^2}{s_2^2}$$

If the two populations are normally distributed and if $H_0 : \sigma_1^2 = \sigma_2^2$ is true then under independent sampling F approximately follows an F-distribution with degrees of freedom $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

A test based on the test statistic F is called an *F-test*.

A most important point is that while the rejection region for a right-tailed test is exactly as in every other situation that we have encountered, because of the asymmetry in the F-distribution the critical value for a left-tailed test and the lower critical value for a two-tailed test have the special forms shown in the following table:

Terminology	Alternative Hypothesis	Rejection Region
Right-tailed	$H_a: \sigma_1^2 > \sigma_2^2$	$F \geq F_\alpha$
Left-tailed	$H_a: \sigma_1^2 < \sigma_2^2$	$F \leq F_{1-\alpha}$
Two-tailed	$H_a: \sigma_1^2 \neq \sigma_2^2$	$F \leq F_{1-\alpha/2}$ or $F \geq F_{\alpha/2}$

Figure **12.1.3** illustrates these rejection regions.

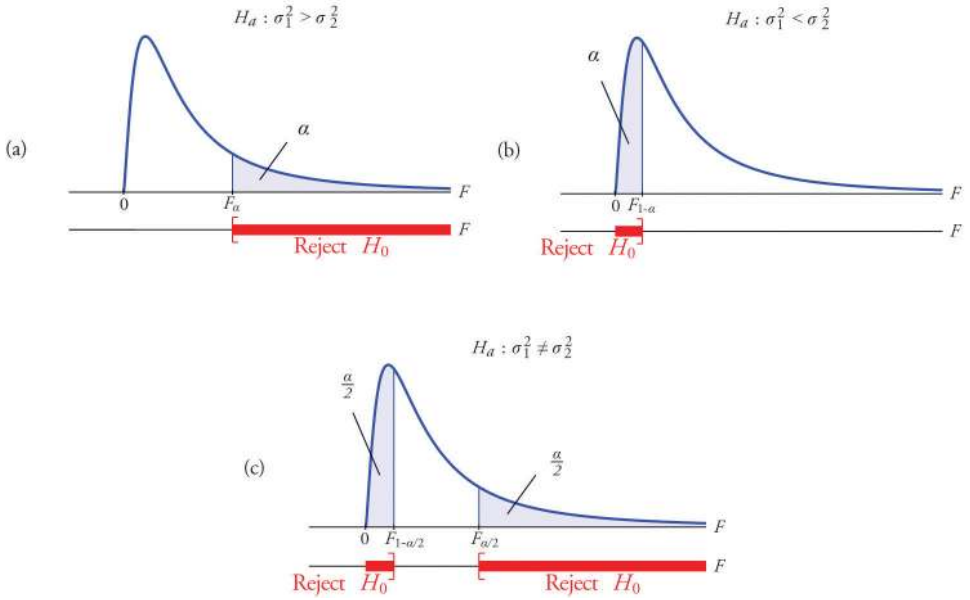


Figure **12.1.3**: Rejection Regions: (a) Right-Tailed; (b) Left-Tailed; (c) Two-Tailed

The test is performed using the usual five-step procedure described at the end of Section 8.1.

Example 12.1.4

One of the quality measures of blood glucose meter strips is the consistency of the test results on the same sample of blood. The consistency is measured by the variance of the readings in repeated testing. Suppose two types of strips, A and B, are compared for their respective consistencies. We arbitrarily label the population of Type A strips Population 1 and the population of Type B strips Population 2. Suppose 15 Type A strips were tested with blood drops from a well-shaken vial and 20 Type B strips were tested with the blood from the same vial. The results are summarized in Table **12.1.3**. Assume the glucose readings using Type A strips follow a normal distribution with variance $\sigma_1^2$and those using Type B strips follow a normal distribution with variance with σ_2^2. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the consistencies of the two types of strips are different.

Table **12.1.3**: Two Types of Test Strips

Strip Type	Sample Size	Sample Variance
A	$n_1 = 16$	$s_1^2 = 2.09$
B	$n_2 = 21$	$s_2^2 = 1.10$

Solution

- **Step 1**. The test of hypotheses is

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$vs.$$
$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ @ } \alpha = 0.10$$

- **Step 2**. The distribution is the F-distribution with degrees of freedom $df_1 = 16 - 1 = 15$ and $df_2 = 21 - 1 = 20$.
- **Step 3**. The test is two-tailed. The left or lower critical value is $F_{1-\alpha} = F_{0.95} = 0.43$. The right or upper critical value is $F_{\alpha/2} = F_{0.05} = 2.20$. Thus the rejection region is $[0, -0.43] \cup [2.20, \infty)$, as illustrated in Figure 12.1.4.

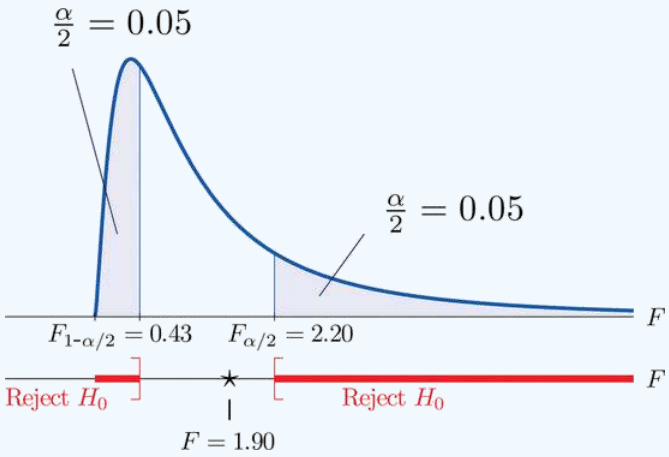


Figure 12.1.4: Rejection Region and Test Statistic for "Example 12.1.4"

- **Step 4**. The value of the test statistic is

$$F = \frac{s_1^2}{s_2^2} = \frac{2.09}{1.10} = 1.90$$

- **Step 5**. As shown in Figure 12.1.4, the test statistic 1.90 does not lie in the rejection region, so the decision is not to reject H_0. The data do not provide sufficient evidence, at the 10% level of significance, to conclude that there is a difference in the consistency, as measured by the variance, of the two types of test strips.

Example 12.1.5

In the context of "Example 12.1.4", suppose Type A test strips are the current market leader and Type B test strips are a newly improved version of Type A. Test, at the 10% level of significance, whether the data given in Table 12.1.3 provide sufficient evidence to conclude that Type B test strips have better consistency (lower variance) than Type A test strips.

Solution

- **Step 1**. The test of hypotheses is now

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$vs.$$
$$H_a : \sigma_1^2 > \sigma_2^2 \text{ @ } \alpha = 0.10$$

- **Step 2**. The distribution is the F-distribution with degrees of freedom $df_1 = 16 - 1 = 15$ and $df_2 = 21 - 1 = 20$.
- **Step 3**. The value of the test statistic is

$$F = \frac{s_1^2}{s_2^2} = \frac{2.09}{1.10} = 1.90$$

- **Step 4**. The test is right-tailed. The single critical value is $F_\alpha = F_{0.10} = 1.84$. Thus the rejection region is $[1.84, \infty)$, as illustrated in Figure 12.1.5.

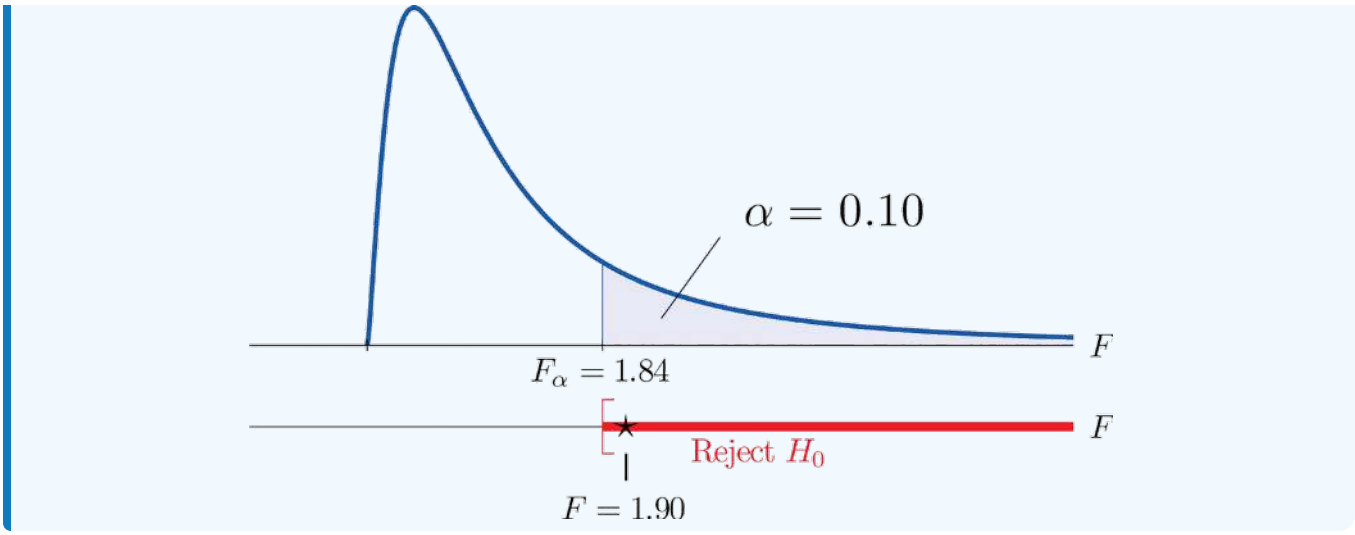
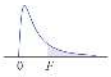


***Figure* 12.1.5**: *Rejection Region and Test Statistic for "Example* 12.1.5*"*

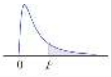
- **Step 5**. As shown in Figure **12.1.5**, the test statistic **1.90** lies in the rejection region, so the decision is to reject H_0 The data provide sufficient evidence, at the **10%** level of significance, to conclude that Type ***B*** test strips have better consistency (lower variance) than Type ***A*** test strips do.

Upper Critical Values of F-Distributions



Upper Critical Values of *F*-Distributions

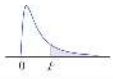
F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.005	1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24630	24836	25044	25253
0.01	1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6157	6209	6261	6313
0.025	1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1010
0.05	1	161	200	216	225	230	234	237	239	241	242	246	248	250	252
0.10	1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	61.2	61.7	62.3	62.8
0.005	2	199	199	199	199	199	199	199	199	199	199	199	199	199	199
0.01	2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
0.025	2	38.5	39.0	39.2	39.3	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
0.05	2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
0.10	2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.46	9.47
0.005	3	55.6	49.8	47.5	46.2	45.4	44.9	44.4	44.1	43.9	43.7	43.1	42.8	42.5	42.2
0.01	3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.4	27.2	26.9	26.7	26.5	26.3
0.025	3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.2	14.1	14.0
0.05	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.57
0.10	3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.15
0.005	4	31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.4	20.2	19.9	19.6
0.01	4	21.2	18.0	16.8	16.0	15.5	15.2	15.0	14.8	14.7	14.6	14.2	14.0	13.9	13.7
0.025	4	12.2	10.7	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.36
0.05	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.69
0.10	4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.82	3.79
0.005	5	22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.2	12.9	12.7	12.4
0.01	5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.72	9.55	9.38	9.20
0.025	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.12
0.05	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.43
0.10	5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24	3.21	3.17	3.14
0.005	6	18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.3	9.81	9.59	9.36	9.12
0.01	6	13.8	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.06
0.025	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.07	4.96
0.05	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.74
0.10	6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.80	2.76
0.005	7	16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	7.97	7.75	7.53	7.31
0.01	7	12.3	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.82
0.025	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.25
0.05	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.30
0.10	7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.56	2.51



Upper Critical Values of *F*-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.005	8	14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	6.81	6.61	6.40	6.18
0.01	8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20	5.03
0.025	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.78
0.05	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.01
0.10	8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.38	2.34
0.005	9	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.03	5.83	5.62	5.41
0.01	9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.48
0.025	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.56	3.45
0.05	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.79
0.10	9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.25	2.21
0.005	10	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.47	5.27	5.07	4.86
0.01	10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25	4.08
0.025	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.20
0.05	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.62
0.10	10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.16	2.11
0.005	11	12.2	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.05	4.86	4.65	4.45
0.01	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	3.94	3.78
0.025	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.12	3.00
0.05	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.49
0.10	11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.12	2.08	2.03
0.005	12	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.72	4.53	4.33	4.12
0.01	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.54
0.025	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	2.96	2.85
0.05	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.38
0.10	12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10	2.06	2.01	1.96
0.005	13	11.4	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.46	4.27	4.07	3.87
0.01	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.51	3.34
0.025	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.84	2.72
0.05	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.30
0.10	13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	2.01	1.96	1.90
0.005	14	11.1	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.25	4.06	3.86	3.66
0.01	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.18
0.025	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.73	2.61
0.05	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.22
0.10	14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.96	1.91	1.86

https://stats.libretexts.org/@go/page/48613

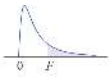


Upper Critical Values of *F*-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.005	15	10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.07	3.88	3.69	3.48
0.01	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.21	3.05
0.025	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.52
0.05	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.16
0.10	15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.87	1.82
0.005	20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.50	3.32	3.12	2.92
0.01	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.61
0.025	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.22
0.05	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.95
0.10	20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.79	1.74	1.68
0.005	30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.01	2.82	2.63	2.42
0.01	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.21
0.025	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	1.94
0.05	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.74
0.10	30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.67	1.61	1.54
0.005	40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.78	2.60	2.40	2.18
0.01	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.02
0.025	40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.94	1.80
0.05	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.64
0.10	40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66	1.61	1.54	1.47
0.005	50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.65	2.47	2.27	2.05
0.01	50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42	2.27	2.10	1.91
0.025	50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.99	1.87	1.72
0.05	50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.58
0.10	50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63	1.57	1.50	1.42
0.005	60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.57	2.39	2.19	1.96
0.01	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.84
0.025	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.67
0.05	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.53
0.10	60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.60	1.54	1.48	1.40
0.005	100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.41	2.23	2.02	1.79
0.01	100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.89	1.69
0.025	100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.85	1.71	1.56
0.05	100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.57	1.45
0.10	100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56	1.49	1.42	1.34

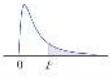
https://stats.libretexts.org/@go/page/48613

Lower Critical Values of F-Distributions



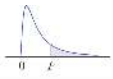
Lower Critical Values of F-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.90	1	0.03	0.12	0.18	0.22	0.25	0.26	0.28	0.29	0.30	0.30	0.33	0.34	0.35	0.36
0.95	1	0.01	0.05	0.10	0.13	0.15	0.17	0.18	0.19	0.20	0.20	0.22	0.23	0.24	0.25
0.975	1	0.00	0.03	0.06	0.08	0.10	0.11	0.12	0.13	0.14	0.14	0.16	0.17	0.18	0.19
0.99	1	0.00	0.01	0.03	0.05	0.06	0.07	0.08	0.09	0.09	0.10	0.12	0.12	0.13	0.14
0.995	1	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12
0.90	2	0.02	0.11	0.18	0.23	0.26	0.29	0.31	0.32	0.33	0.34	0.37	0.39	0.40	0.42
0.95	2	0.01	0.05	0.10	0.14	0.17	0.19	0.21	0.22	0.23	0.24	0.27	0.29	0.30	0.32
0.975	2	0.00	0.03	0.06	0.09	0.12	0.14	0.15	0.17	0.17	0.18	0.21	0.22	0.24	0.25
0.99	2	0.00	0.01	0.03	0.06	0.08	0.09	0.10	0.12	0.12	0.13	0.16	0.17	0.19	0.20
0.995	2	0.00	0.01	0.02	0.04	0.05	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.16	0.17
0.90	3	0.02	0.11	0.19	0.24	0.28	0.30	0.33	0.34	0.36	0.37	0.40	0.42	0.44	0.46
0.95	3	0.00	0.05	0.11	0.15	0.18	0.21	0.23	0.25	0.26	0.27	0.30	0.32	0.34	0.36
0.975	3	0.00	0.03	0.06	0.10	0.13	0.15	0.17	0.18	0.20	0.21	0.24	0.26	0.28	0.30
0.99	3	0.00	0.01	0.03	0.06	0.08	0.10	0.12	0.13	0.14	0.15	0.18	0.20	0.22	0.24
0.995	3	0.00	0.01	0.02	0.04	0.06	0.08	0.09	0.10	0.11	0.12	0.15	0.17	0.19	0.21
0.90	4	0.02	0.11	0.19	0.24	0.28	0.31	0.34	0.36	0.37	0.38	0.42	0.44	0.47	0.49
0.95	4	0.00	0.05	0.11	0.16	0.19	0.22	0.24	0.26	0.28	0.29	0.33	0.35	0.37	0.40
0.975	4	0.00	0.03	0.07	0.10	0.14	0.16	0.18	0.20	0.21	0.22	0.26	0.28	0.31	0.33
0.99	4	0.00	0.01	0.03	0.06	0.09	0.11	0.13	0.14	0.16	0.17	0.20	0.23	0.25	0.27
0.995	4	0.00	0.01	0.02	0.04	0.06	0.08	0.10	0.11	0.13	0.14	0.17	0.19	0.22	0.24
0.90	5	0.02	0.11	0.19	0.25	0.29	0.32	0.35	0.37	0.38	0.40	0.44	0.46	0.49	0.51
0.95	5	0.00	0.05	0.11	0.16	0.20	0.23	0.25	0.27	0.29	0.30	0.34	0.37	0.39	0.42
0.975	5	0.00	0.03	0.07	0.11	0.14	0.17	0.19	0.21	0.22	0.24	0.28	0.30	0.33	0.36
0.99	5	0.00	0.01	0.04	0.06	0.09	0.11	0.13	0.15	0.17	0.18	0.22	0.24	0.27	0.30
0.995	5	0.00	0.01	0.02	0.04	0.07	0.09	0.11	0.12	0.13	0.15	0.19	0.21	0.24	0.27
0.90	6	0.02	0.11	0.19	0.25	0.29	0.33	0.35	0.37	0.39	0.41	0.45	0.48	0.50	0.53
0.95	6	0.00	0.05	0.11	0.16	0.20	0.23	0.26	0.28	0.30	0.31	0.36	0.38	0.41	0.44
0.975	6	0.00	0.03	0.07	0.11	0.14	0.17	0.20	0.21	0.23	0.25	0.29	0.32	0.35	0.38
0.99	6	0.00	0.01	0.04	0.07	0.09	0.12	0.14	0.16	0.17	0.19	0.23	0.26	0.29	0.32
0.995	6	0.00	0.01	0.02	0.05	0.07	0.09	0.11	0.13	0.14	0.15	0.2	00.22	0.25	0.29
0.90	7	0.02	0.11	0.19	0.25	0.30	0.33	0.36	0.38	0.40	0.41	0.46	0.49	0.52	0.55
0.95	7	0.00	0.05	0.11	0.16	0.21	0.24	0.26	0.29	0.30	0.32	0.37	0.40	0.43	0.46
0.975	7	0.00	0.03	0.07	0.11	0.15	0.18	0.20	0.22	0.24	0.25	0.30	0.33	0.36	0.40
0.99	7	0.00	0.01	0.04	0.07	0.10	0.12	0.14	0.16	0.18	0.19	0.24	0.27	0.30	0.34
0.995	7	0.00	0.01	0.02	0.05	0.07	0.09	0.11	0.13	0.15	0.16	0.21	0.23	0.27	0.3



Lower Critical Values of F-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.90	8	0.02	0.11	0.19	0.25	0.30	0.34	0.36	0.39	0.40	0.42	0.47	0.50	0.53	0.56
0.95	8	0.00	0.05	0.11	0.17	0.21	0.24	0.27	0.29	0.31	0.33	0.38	0.41	0.44	0.48
0.975	8	0.00	0.03	0.07	0.11	0.15	0.18	0.20	0.23	0.24	0.26	0.31	0.34	0.38	0.41
0.99	8	0.00	0.01	0.04	0.07	0.10	0.12	0.15	0.17	0.18	0.20	0.25	0.28	0.32	0.35
0.995	8	0.00	0.01	0.02	0.05	0.07	0.09	0.12	0.13	0.15	0.16	0.21	0.24	0.28	0.32
0.90	9	0.02	0.11	0.19	0.25	0.30	0.34	0.37	0.39	0.41	0.43	0.48	0.51	0.54	0.58
0.95	9	0.00	0.05	0.11	0.17	0.21	0.24	0.27	0.30	0.31	0.33	0.39	0.42	0.45	0.49
0.975	9	0.00	0.03	0.07	0.11	0.15	0.18	0.21	0.23	0.25	0.26	0.32	0.35	0.39	0.43
0.99	9	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.17	0.19	0.20	0.26	0.29	0.33	0.37
0.995	9	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.15	0.17	0.22	0.25	0.29	0.33
0.90	10	0.02	0.11	0.19	0.26	0.30	0.34	0.37	0.39	0.41	0.43	0.49	0.52	0.55	0.59
0.95	10	0.00	0.05	0.11	0.17	0.21	0.25	0.27	0.30	0.32	0.34	0.39	0.43	0.46	0.50
0.975	10	0.00	0.03	0.07	0.11	0.15	0.18	0.21	0.23	0.25	0.27	0.33	0.36	0.40	0.44
0.99	10	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.17	0.19	0.21	0.26	0.30	0.34	0.38
0.995	10	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.16	0.17	0.23	0.26	0.30	0.34
0.90	11	0.02	0.11	0.19	0.26	0.30	0.34	0.37	0.40	0.42	0.43	0.49	0.52	0.56	0.60
0.95	11	0.00	0.05	0.11	0.17	0.21	0.25	0.28	0.30	0.32	0.34	0.40	0.43	0.47	0.51
0.975	11	0.00	0.03	0.07	0.11	0.15	0.18	0.21	0.24	0.26	0.27	0.33	0.37	0.41	0.45
0.99	11	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.17	0.19	0.21	0.27	0.30	0.34	0.39
0.995	11	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.16	0.17	0.23	0.27	0.31	0.36
0.90	12	0.02	0.11	0.19	0.26	0.31	0.34	0.37	0.40	0.42	0.44	0.50	0.53	0.56	0.60
0.95	12	0.00	0.05	0.11	0.17	0.21	0.25	0.28	0.30	0.33	0.34	0.40	0.44	0.48	0.52
0.975	12	0.00	0.03	0.07	0.11	0.15	0.19	0.21	0.24	0.26	0.28	0.34	0.37	0.41	0.46
0.99	12	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.18	0.20	0.21	0.27	0.31	0.35	0.40
0.995	12	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.16	0.18	0.24	0.27	0.31	0.36
0.90	13	0.02	0.11	0.19	0.26	0.31	0.35	0.38	0.40	0.42	0.44	0.50	0.53	0.57	0.61
0.95	13	0.00	0.05	0.11	0.17	0.21	0.25	0.28	0.31	0.33	0.35	0.41	0.44	0.48	0.53
0.975	13	0.00	0.03	0.07	0.11	0.15	0.19	0.22	0.24	0.26	0.28	0.34	0.38	0.42	0.47
0.99	13	0.00	0.01	0.04	0.07	0.10	0.13	0.16	0.18	0.20	0.22	0.28	0.31	0.36	0.41
0.995	13	0.00	0.01	0.02	0.05	0.08	0.10	0.12	0.14	0.16	0.18	0.24	0.28	0.32	0.37
0.90	14	0.02	0.11	0.19	0.26	0.31	0.35	0.38	0.40	0.43	0.44	0.50	0.54	0.58	0.62
0.95	14	0.00	0.05	0.11	0.17	0.22	0.25	0.28	0.31	0.33	0.35	0.41	0.45	0.49	0.54
0.975	14	0.00	0.03	0.07	0.12	0.15	0.19	0.22	0.24	0.26	0.28	0.35	0.38	0.43	0.48
0.99	14	0.00	0.01	0.04	0.07	0.10	0.13	0.16	0.18	0.20	0.22	0.28	0.32	0.36	0.42
0.995	14	0.00	0.01	0.02	0.05	0.08	0.10	0.12	0.15	0.16	0.18	0.24	0.28	0.33	0.38



Lower Critical Values of *F*-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.90	15	0.02	0.11	0.19	0.26	0.31	0.35	0.38	0.41	0.43	0.45	0.51	0.54	0.58	0.62
0.95	15	0.00	0.05	0.11	0.17	0.22	0.25	0.28	0.31	0.33	0.35	0.42	0.46	0.50	0.54
0.975	15	0.00	0.03	0.07	0.12	0.16	0.19	0.22	0.24	0.27	0.28	0.35	0.39	0.43	0.49
0.99	15	0.00	0.01	0.04	0.07	0.10	0.13	0.16	0.18	0.20	0.22	0.28	0.32	0.37	0.43
0.995	15	0.00	0.01	0.02	0.05	0.08	0.10	0.13	0.15	0.17	0.18	0.25	0.29	0.33	0.39
0.90	20	0.02	0.11	0.19	0.26	0.31	0.35	0.39	0.41	0.44	0.45	0.52	0.56	0.60	0.65
0.95	20	0.00	0.05	0.12	0.17	0.22	0.26	0.29	0.32	0.34	0.36	0.43	0.47	0.52	0.57
0.975	20	0.00	0.03	0.07	0.12	0.16	0.19	0.22	0.25	0.27	0.29	0.36	0.41	0.46	0.51
0.99	20	0.00	0.01	0.04	0.07	0.10	0.14	0.16	0.19	0.21	0.23	0.30	0.34	0.39	0.45
0.995	20	0.00	0.01	0.02	0.05	0.08	0.1	00.13	0.15	0.17	0.19	0.26	0.30	0.35	0.42
0.90	30	0.02	0.11	0.19	0.26	0.32	0.36	0.39	0.42	0.44	0.46	0.53	0.58	0.62	0.68
0.95	30	0.00	0.05	0.12	0.17	0.22	0.26	0.30	0.32	0.35	0.37	0.45	0.49	0.54	0.61
0.975	30	0.00	0.03	0.07	0.12	0.16	0.20	0.23	0.26	0.28	0.30	0.38	0.43	0.48	0.55
0.99	30	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.19	0.22	0.24	0.31	0.36	0.42	0.49
0.995	30	0.00	0.01	0.02	0.05	0.08	0.11	0.13	0.16	0.18	0.20	0.27	0.32	0.38	0.46
0.90	40	0.02	0.11	0.19	0.26	0.32	0.36	0.39	0.42	0.45	0.47	0.54	0.59	0.64	0.70
0.95	40	0.00	0.05	0.12	0.17	0.22	0.26	0.30	0.33	0.35	0.38	0.45	0.50	0.56	0.63
0.975	40	0.00	0.03	0.07	0.12	0.16	0.20	0.23	0.26	0.29	0.31	0.39	0.44	0.50	0.57
0.99	40	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.22	0.24	0.32	0.37	0.43	0.52
0.995	40	0.00	0.01	0.02	0.05	0.08	0.11	0.13	0.16	0.18	0.20	0.28	0.33	0.40	0.48
0.90	50	0.02	0.11	0.19	0.26	0.32	0.36	0.40	0.43	0.45	0.47	0.55	0.59	0.64	0.71
0.95	50	0.00	0.05	0.12	0.18	0.23	0.27	0.30	0.33	0.36	0.38	0.46	0.51	0.57	0.64
0.975	50	0.00	0.03	0.07	0.12	0.16	0.20	0.23	0.26	0.29	0.31	0.39	0.44	0.51	0.59
0.99	50	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.22	0.24	0.32	0.38	0.45	0.53
0.995	50	0.00	0.01	0.02	0.05	0.08	0.11	0.14	0.16	0.18	0.20	0.28	0.34	0.41	0.50
0.90	60	0.02	0.11	0.19	0.26	0.32	0.36	0.40	0.43	0.45	0.47	0.55	0.60	0.65	0.72
0.95	60	0.00	0.05	0.12	0.18	0.23	0.27	0.30	0.33	0.36	0.38	0.46	0.51	0.57	0.65
0.975	60	0.00	0.03	0.07	0.12	0.16	0.20	0.24	0.26	0.29	0.31	0.40	0.45	0.52	0.60
0.99	60	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.22	0.24	0.33	0.38	0.45	0.54
0.995	60	0.00	0.01	0.02	0.05	0.08	0.11	0.14	0.16	0.18	0.21	0.29	0.34	0.41	0.51
0.90	100	0.02	0.11	0.19	0.26	0.32	0.36	0.40	0.43	0.46	0.48	0.56	0.61	0.66	0.74
0.95	100	0.00	0.05	0.12	0.18	0.23	0.27	0.31	0.34	0.36	0.39	0.47	0.52	0.59	0.68
0.975	100	0.00	0.03	0.07	0.12	0.16	0.20	0.24	0.27	0.29	0.32	0.40	0.46	0.53	0.63
0.99	100	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.23	0.25	0.34	0.39	0.47	0.57
0.995	100	0.00	0.01	0.02	0.05	0.08	0.11	0.14	0.16	0.19	0.21	0.29	0.35	0.43	0.54

Key Takeaway

- Critical values of an F-distribution with degrees of freedom df_1 and df_2 are found in tables above.
- An F-test can be used to evaluate the hypothesis of two identical normal population variances.

This page titled 12.1: F-Tests is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **11.3: F-tests for Equality of Two Variances** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

12.2: F-Tests in One-Way ANOVA

Learning Objectives

- To understand how to use an F-test to judge whether several population means are all equal

In Chapter 9, we saw how to compare two population means μ_1 and μ_2.

In this section we will learn to compare three or more population means at the same time, which is often of interest in practical applications. For example, an administrator at a university may be interested in knowing whether student grade point averages are the same for different majors. In another example, an oncologist may be interested in knowing whether patients with the same type of cancer have the same average survival times under several different competing cancer treatments.

In general, suppose there are K normal populations with possibly different means, $\mu_1, \mu_2, \ldots, \mu_K$, but all with the same variance σ^2. The study question is whether all the K population means are the same. We formulate this question as the test of hypotheses

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$
$$vs.$$
$$H_a : \text{not all K population means are equal}$$

To perform the test K independent random samples are taken from the K normal populations. The K sample means, the K sample variances, and the K sample sizes are summarized in the table:

Population	Sample Size	Sample Mean	Sample Variance
1	n_1	$\bar{x}_1$	s_1^2
2	n_2	$\bar{x}_2$	s_2^2
⋮	⋮	⋮	⋮
K	n_K	$\bar{x}_K$	s_K^2

Define the following quantities:

Definitions

The *combined sample size*:

$$n = n_1 + n_2 + \ldots + n_K$$

The *mean of the combined sample* of all n observations:

$$x = \frac{\Sigma x}{n} = \frac{n_1\bar{x} + n_2\bar{x}_2 + \ \ldots \ + n_K\bar{x}_K}{n}$$

The *mean square for treatment:*

$$MST = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots \ + n_K(\bar{x}K - \bar{x})^2}{K - 1}$$

The *mean square for error:*

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ \ldots \ + (n_{K-1})s_K^2}{n - K}$$

MST can be thought of as the variance between the K individual independent random samples and MSE as the variance within the samples. This is the reason for the name “analysis of variance,” universally abbreviated ANOVA. The adjective “one-way” has to do with the fact that the sampling scheme is the simplest possible, that of taking one random sample from each population under consideration. If the means of the K populations are all the same then the two quantities MST and MSE should be close to the

same, so the null hypothesis will be rejected if the ratio of these two quantities is significantly greater than 1. This yields the following test statistic and methods and conditions for its use.

Test Statistic for Testing the Null Hypothesis that K Population Means Are Equal

$$F = \frac{MST}{MSE}$$

If the K populations are normally distributed with a common variance and if $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ is true then under independent random sampling F approximately follows an F-distribution with degrees of freedom $df_1 = K - 1$ and $df_2 = n - K$.

The test is right-tailed: H_0 is rejected at level of significance α if $F \geq F_\alpha$.

As always the test is performed using the usual five-step procedure.

Example 12.2.1

The average of grade point averages (GPAs) of college courses in a specific major is a measure of difficulty of the major. An educator wishes to conduct a study to find out whether the difficulty levels of different majors are the same. For such a study, a random sample of major grade point averages (GPA) of 11 graduating seniors at a large university is selected for each of the four majors mathematics, English, education, and biology. The data are given in Table 12.2.1. Test, at the 5% level of significance, whether the data contain sufficient evidence to conclude that there are differences among the average major GPAs of these four majors.

Table 12.2.1: Difficulty Levels of College Majors

Mathematics	English	Education	Biology
2.59	3.64	4.00	2.78
3.13	3.19	3.59	3.51
2.97	3.15	2.80	2.65
2.50	3.78	2.39	3.16
2.53	3.03	3.47	2.94
3.29	2.61	3.59	2.32
2.53	3.20	3.74	2.58
3.17	3.30	3.77	3.21
2.70	3.54	3.13	3.23
3.88	3.25	3.00	3.57
2.64	4.00	3.47	3.22

Solution

- **Step 1**. The test of hypotheses is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$vs.$$
$$H_a : \text{not all four population means are equal @ } \alpha = 0.05$$

- **Step 2**. The test statistic is $F = MST/MSE$ with (since $n = 44$ and $K = 4$) degrees of freedom $df_1 = K - 1 = 4 - 1 = 3$ and $df_2 = n - K = 44 - 4 = 40$.

- **Step 3**. If we index the population of mathematics majors by **1**, English majors by **2**, education majors by **3**, and biology majors by **4**, then the sample sizes, sample means, and sample variances of the four samples in Table **12.2.1** are summarized (after rounding for simplicity) by:

Major	Sample Size	Sample Mean	Sample Variance
Mathematics	$n_1 = 11$	$\bar{x}_1 = 2.90$	$s_1^2 = 0.188$
English	$n_2 = 11$	$\bar{x}_2 = 3.34$	$s_2^2 = 0.148$
Education	$n_3 = 11$	$\bar{x}_3 = 3.36$	$s_3^2 = 0.229$
Biology	$n_4 = 11$	$\bar{x}_4 = 3.02$	$s_4^2 = 0.157$

The average of all **44** observations is (after rounding for simplicity) $\bar{x} = 3.15$. We compute (rounding for simplicity)

$$\begin{aligned} MST &= \frac{11(2.90-3.15)^2 + 11(3.34-3.15)^2 + 11(3.36-3.15)^2 + 11(3.02-3.15)^2}{4-1} \\ &= \frac{1.7556}{3} \\ &= 0.585 \end{aligned}$$

and

$$\begin{aligned} MSE &= \frac{(11-1)(0.188) + (11-1)(0.148) + (11-1)(0.229) + (11-1)(0.157)}{44-4} \\ &= \frac{7.22}{40} \\ &= 0.181 \end{aligned}$$

so that

$$F = \frac{MST}{MSE} = \frac{0.585}{0.181} = 3.232$$

- **Step 4**. The test is right-tailed. The single critical value is (since $df_1 = 3$ and $df_2 = 40$) $F_\alpha = F_{0.05} = 2.84$. Thus the rejection region is $[2.84, \infty)$, as illustrated in Figure **12.2.1**.

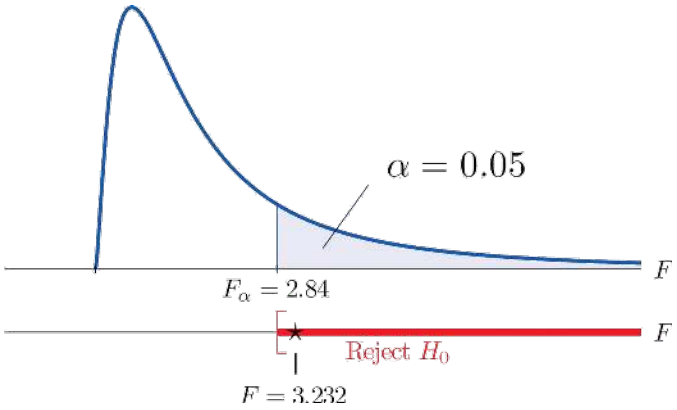


Figure **12.2.1**: Rejection Region

- **Step 5**. Since $F = 3.232 > 2.84$, we reject H_0. The data provide sufficient evidence, at the **5%** level of significance, to conclude that the averages of major GPAs for the four majors considered are not all equal.

Example 12.2.2: Mice Survival Times

A research laboratory developed two treatments which are believed to have the potential of prolonging the survival times of patients with an acute form of thymic leukemia. To evaluate the potential treatment effects **33** laboratory mice with thymic

leukemia were randomly divided into three groups. One group received Treatment 1, one received Treatment 2, and the third was observed as a control group. The survival times of these mice are given in Table 12.2.2. Test, at the 1% level of significance, whether these data provide sufficient evidence to confirm the belief that at least one of the two treatments affects the average survival time of mice with thymic leukemia.

Table 12.2.2 Mice Survival Times in Days

Treatment 1		Treatment 2	Control
71	75	77	81
72	73	67	79
75	72	79	73
80	65	78	71
60	63	81	75
65	69	72	84
63	64	71	77
78	71	84	67
		91	

Solution

- **Step 1**. The test of hypotheses is

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$vs.$$
$$H_a : \text{not all three population means are equal @ } \alpha = 0.01$$

- **Step 2**. The test statistic is $F = \dfrac{MST}{MSE}$ with (since $n = 33$ and $K = 3$) degrees of freedom $df_1 = K - 1 = 3 - 1 = 2$ and $df_2 = n - K = 33 - 3 = 30$.
- **Step 3**. If we index the population of mice receiving Treatment 1 by 1, Treatment 2 by 2, and no treatment by 3, then the sample sizes, sample means, and sample variances of the three samples in Table 12.2.2 are summarized (after rounding for simplicity) by:

Table 12.2.2: Mice Survival Times in Days

Group	Sample Size	Sample Mean	Sample Variance
Treatment 1	$n_1 = 16$	$\bar{x}_1 = 69.75$	$s_1^2 = 34.47$
Treatment 2	$n_2 = 9$	$\bar{x}_2 = 77.78$	$s_2^2 = 52.69$
Control	$n_3 = 8$	$\bar{x}_3 = 75.88$	$s_3^2 = 30.69$

The average of all 33 observations is (after rounding for simplicity) $\bar{x} = 73.42$. We compute (rounding for simplicity)

$$\begin{aligned} MST &= \frac{16(69.75 - 73.42)^2 + 9(77.78 - 73.42)^2 + 8(75.88 - 73.42)^2}{31} \\ &= \frac{434.63}{2} \\ &= 217.50 \end{aligned}$$

and

$$MSE = \frac{(16-1)(34.47)+(9-1)(52.69)+(8-1)(30.69)}{33-3}$$
$$= \frac{1153.4}{30}$$
$$= 38.45$$

so that

$$F = \frac{MST}{MSE} = \frac{217.50}{38.45} = 5.65$$

- **Step 4**. The test is right-tailed. The single critical value is $F_\alpha = F_{0.01} = 5.39$. Thus the rejection region is $[5.39, \infty)$, as illustrated in Figure 12.2.2.

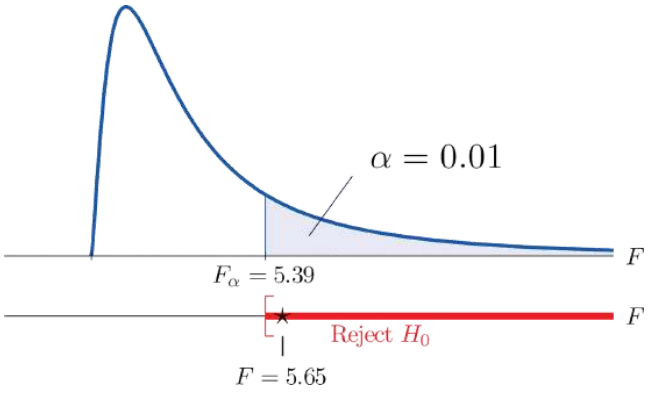


Figure 12.2.2: Rejection Region

- **Step 5**. Since $F = 5.65 > 5.39$, we reject H_0. The data provide sufficient evidence, at the 1% level of significance, to conclude that a treatment effect exists at least for one of the two treatments in increasing the mean survival time of mice with thymic leukemia.

It is important to to note that, if the null hypothesis of equal population means is rejected, the statistical implication is that not all population means are equal. It does not however tell which population mean is different from which. The inference about where the suggested difference lies is most frequently made by a follow-up study.

Key Takeaway

- An F-test can be used to evaluate the hypothesis that the means of several normal populations, all with the same standard deviation, are identical.

This page titled 12.2: F-Tests in One-Way ANOVA is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **11.4: F-Tests in One-Way ANOVA** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

Index

A

Adding probabilities
4.2: Addition and Multiplication Rule of Probability
alternative hypothesis
8.1: The Elements of Hypothesis Testing
anova
12.2: F-Tests in One-Way ANOVA

B

bar graph
2.3: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
binomial probability distribution
5.2: The Binomial Distribution
binomial random variable
5.2: The Binomial Distribution
blinding
1.3.1: Experimental Design and Ethics
box plots
3.3: Relative Position of Data

C

Chebyshev's Theorem
3.4: The Empirical Rule and Chebyshev's Theorem
cluster sampling
1.2: Data, Sampling, and Variation in Data and Sampling
2.1: Data, Sampling, and Variation in Data and Sampling
coefficient of determination
10.2: The Regression Equation and Correlation Coefficient
combined sample size
12.2: F-Tests in One-Way ANOVA
Comparing two population means
9.2.1: Large, Independent Samples
9.2.2: Small, Independent Samples
Comparing Two Population Proportions
9.1: Two Population Proportions
complement
4.1.1: Terminology
4.1.2: Independent and Mutually Exclusive Events
conditional probability
4.1.1: Terminology
confidence interval for estimating a population mean
7.3: Sample Size Considerations
CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS
9.1: Two Population Proportions
confidence interval for the difference in two population means
9.4: Sample Size Considerations
confidence interval formula for estimating a population proportion
7.3: Sample Size Considerations
Confidence Intervals for a Proportion
7.1: Estimation of a Population Proportion
7.3: Sample Size Considerations
contingency table
4.3: Conditional Probability using Contingency Tables
11.1: Chi-Square Tests for Independence
continuous data
1.2: Data, Sampling, and Variation in Data and Sampling
2.1: Data, Sampling, and Variation in Data and Sampling
control group
1.3.1: Experimental Design and Ethics
critical value test
8.3: Tests for a Population Proportion
Cumulative Normal Probability
6.1.2: The Standard Normal Distribution
cumulative probability distributions
5.2: The Binomial Distribution
cumulative relative frequency
1.3: Frequency, Frequency Tables, and Levels of Measurement

D

DENSITY FUNCTION
6.1.1: Continuous Random Variables
direction of a relationship between the variables
10.1.2: Scatter Plots
discrete data
1.2: Data, Sampling, and Variation in Data and Sampling
2.1: Data, Sampling, and Variation in Data and Sampling

E

empirical rule
3.4: The Empirical Rule and Chebyshev's Theorem
7.2.1: Large Sample Estimation of a Population Mean
Equal variance
10.3: Testing for Significance Linear Correlation
ethics
1.3.1: Experimental Design and Ethics
event
4.1.1: Terminology
experimental unit
1.3.1: Experimental Design and Ethics
explanatory variable
1.3.1: Experimental Design and Ethics
extrapolation
10.4: Prediction

F

frequency
1.3: Frequency, Frequency Tables, and Levels of Measurement
frequency table
1.3: Frequency, Frequency Tables, and Levels of Measurement

H

histograms
2.2: Histogram
Hypothesis Testing
8.1: The Elements of Hypothesis Testing

I

independent events
4.1.2: Independent and Mutually Exclusive Events
4.2: Addition and Multiplication Rule of Probability
Institutional Review Board
1.3.1: Experimental Design and Ethics
interpolation
10.4: Prediction

L

level of measurement
1.3: Frequency, Frequency Tables, and Levels of Measurement
level of significance
8.1: The Elements of Hypothesis Testing
line graph
2.3: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
linear correlation coefficient
10.2: The Regression Equation and Correlation Coefficient
10.3: Testing for Significance Linear Correlation
linear equations
10.1.1: Linear Equations
LINEAR REGRESSION MODEL
10.2: The Regression Equation and Correlation Coefficient
lurking variable
1.3.1: Experimental Design and Ethics

M

margin of error
7.2.1: Large Sample Estimation of a Population Mean
mean
5.1.1: Probability Distributions for Discrete Random Variables
mean square for error
12.2: F-Tests in One-Way ANOVA
mean square for treatment
12.2: F-Tests in One-Way ANOVA
Minimum Sample Size for Estimating a Population Mean
7.3: Sample Size Considerations
mode
3.1: Measures of Center
most conservative estimate
7.3: Sample Size Considerations
Multiplying probabilities
4.2: Addition and Multiplication Rule of Probability
mutually exclusive
4.1.2: Independent and Mutually Exclusive Events
4.2: Addition and Multiplication Rule of Probability

N

normal distribution
6.1.1: Continuous Random Variables
6.3: The Central Limit Theorem for Sample Means
null hypothesis
8.1: The Elements of Hypothesis Testing

O

observed significance
- 8.2.1: The Observed Significance of a Test

outcome
- 4.1.1: Terminology

P

paired difference samples
- 9.3: Two Population Means - Paired Samples

paired samples
- 9.3: Two Population Means - Paired Samples

parameter
- 1.1: Definitions of Statistics and Key Terms

Pareto chart
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 2.1: Data, Sampling, and Variation in Data and Sampling

percentiles
- 3.3: Relative Position of Data

placebo
- 1.3.1: Experimental Design and Ethics

pooled variance
- 9.2.2: Small, Independent Samples

population
- 1.1: Definitions of Statistics and Key Terms

population mean
- 3.1: Measures of Center

population median
- 3.1: Measures of Center

population mode
- 3.1: Measures of Center

prediction
- 10.4: Prediction

probability
- 1.1: Definitions of Statistics and Key Terms

probability distribution function
- 5.1.1: Probability Distributions for Discrete Random Variables

Q

qualitative data
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 2.1: Data, Sampling, and Variation in Data and Sampling

quantitative data
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 2.1: Data, Sampling, and Variation in Data and Sampling

quartiles
- 3.3: Relative Position of Data

R

random assignment
- 1.3.1: Experimental Design and Ethics

range
- 3.2: Measures of Variability

rare events
- 8.2.1: The Observed Significance of a Test

relative frequency histograms
- 2.2: Histogram

response variable
- 1.3.1: Experimental Design and Ethics

rounding
- 1.3: Frequency, Frequency Tables, and Levels of Measurement

S

sample mean
- 3.1: Measures of Center

sample median
- 3.1: Measures of Center

sample mode
- 3.1: Measures of Center

sample size
- 9.4: Sample Size Considerations

sample space
- 4.1.1: Terminology

sampling bias
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 2.1: Data, Sampling, and Variation in Data and Sampling

sampling distribution of the mean
- 6.3: The Central Limit Theorem for Sample Means

Sampling Error
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 2.1: Data, Sampling, and Variation in Data and Sampling

sampling with replacement
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 2.1: Data, Sampling, and Variation in Data and Sampling
- 4.1.2: Independent and Mutually Exclusive Events

sampling without replacement
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 2.1: Data, Sampling, and Variation in Data and Sampling
- 4.1.2: Independent and Mutually Exclusive Events

scatter plot
- 10.1.2: Scatter Plots

Skewed
- 3.1: Measures of Center

slope
- 10.1.1: Linear Equations

standard deviation
- 3.2: Measures of Variability
- 5.1.1: Probability Distributions for Discrete Random Variables

standard error of the mean
- 6.3: The Central Limit Theorem for Sample Means

standard normal random variable
- 6.1.2: The Standard Normal Distribution

statistic
- 1.1: Definitions of Statistics and Key Terms

stemplot
- 2.3: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

strength of a relationship between the variables
- 10.1.2: Scatter Plots

T

Tests for Independence
- 11.1: Chi-Square Tests for Independence

The AND Event
- 4.1.1: Terminology

The Or Event
- 4.1.1: Terminology

The OR of Two Events
- 4.1.2: Independent and Mutually Exclusive Events

treatments
- 1.3.1: Experimental Design and Ethics

Type I error
- 8.1: The Elements of Hypothesis Testing

Type II error
- 8.1: The Elements of Hypothesis Testing

V

variable
- 1.1: Definitions of Statistics and Key Terms

variance
- 3.2: Measures of Variability

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Math 11: Elementary Statistics

Webpages: 85

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 3.0: 47.1% (40 pages)
- CC BY 4.0: 32.9% (28 pages)
- Undeclared: 18.8% (16 pages)
- CC BY-NC-SA 4.0: 1.2% (1 page)

By Page

- Math 11: Elementary Statistics - *CC BY-NC-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Statistics - *CC BY-NC-SA 3.0*
    - 1.1: Definitions of Statistics and Key Terms - *CC BY 4.0*
    - 1.2: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 1.3: Frequency, Frequency Tables, and Levels of Measurement - *CC BY 4.0*
      - 1.3.1: Experimental Design and Ethics - *CC BY 4.0*
  - 2: Data Displays - *Undeclared*
    - 2.1: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 2.2: Histogram - *CC BY-NC-SA 3.0*
    - 2.3: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs - *CC BY 4.0*
  - 3: Descriptive Statistics - *CC BY-NC-SA 3.0*
    - 3.1: Measures of Center - *CC BY-NC-SA 3.0*
    - 3.2: Measures of Variability - *CC BY-NC-SA 3.0*
    - 3.3: Relative Position of Data - *CC BY-NC-SA 3.0*
    - 3.4: The Empirical Rule and Chebyshev's Theorem - *CC BY-NC-SA 3.0*
  - 4: Probability Topics - *CC BY 4.0*
    - 4.1: Introduction - *CC BY 4.0*
      - 4.1.1: Terminology - *CC BY 4.0*
      - 4.1.2: Independent and Mutually Exclusive Events - *CC BY 4.0*
    - 4.2: Addition and Multiplication Rule of Probability - *CC BY 4.0*
    - 4.3: Conditional Probability using Contingency Tables - *CC BY 4.0*
    - 4.E: Probability Topics (Exericses) - *CC BY 4.0*
  - 5: Discrete Random Variables - *CC BY-NC-SA 3.0*
    - 5.1: Random Variables - *CC BY-NC-SA 3.0*
      - 5.1.1: Probability Distributions for Discrete Random Variables - *CC BY-NC-SA 3.0*
    - 5.2: The Binomial Distribution - *CC BY-NC-SA 3.0*
    - 5.E: Discrete Random Variables (Exercises) - *CC BY-NC-SA 3.0*
  - 6: Continuous Random Variables - *CC BY-NC-SA 3.0*
    - 6.1: The Standard Normal Distribution - *Undeclared*
      - 6.1.1: Continuous Random Variables - *CC BY-NC-SA 3.0*
      - 6.1.2: The Standard Normal Distribution - *CC BY-NC-SA 3.0*
    - 6.2: The General Normal Distribution - *CC BY-NC-SA 3.0*
    - 6.3: The Central Limit Theorem for Sample Means - *CC BY 4.0*
      - 6.3E: The Central Limit Theorem for Sample Means (Exercises) - *CC BY 4.0*
  - 7: Estimation - *CC BY-NC-SA 3.0*
    - 7.1: Estimation of a Population Proportion - *CC BY-NC-SA 3.0*
    - 7.2: Estimation of a Population Mean - *Undeclared*
      - 7.2.1: Large Sample Estimation of a Population Mean - *CC BY-NC-SA 3.0*
      - 7.2.2: Small Sample Estimation of a Population Mean - *CC BY-NC-SA 3.0*

- 7.3: Sample Size Considerations - *CC BY-NC-SA 3.0*
- 7.E: Estimation (Exercises) - *CC BY-NC-SA 3.0*

8: Testing Hypotheses - *CC BY-NC-SA 3.0*
- 8.1: The Elements of Hypothesis Testing - *CC BY-NC-SA 3.0*
- 8.2: Tests for a Population Mean - *CC BY-NC-SA 3.0*
  - 8.2.1: The Observed Significance of a Test - *CC BY-NC-SA 3.0*
  - 8.2.2: Small Sample Tests for a Population Mean - *CC BY-NC-SA 3.0*
- 8.3: Tests for a Population Proportion - *CC BY-NC-SA 3.0*
- 8.E: Testing Hypotheses (Exercises) - *CC BY-NC-SA 3.0*

9: Two-Sample Problems - *CC BY-NC-SA 3.0*
- 9.1: Two Population Proportions - *CC BY-NC-SA 3.0*
- 9.2: Two Population Means - Independent Samples - *Undeclared*
  - 9.2.1: Large, Independent Samples - *CC BY-NC-SA 3.0*
  - 9.2.2: Small, Independent Samples - *CC BY-NC-SA 3.0*
- 9.3: Two Population Means - Paired Samples - *CC BY-NC-SA 3.0*
- 9.4: Sample Size Considerations - *CC BY-NC-SA 3.0*
- 9.E: Two-Sample Problems (Exercises) - *CC BY-NC-SA 3.0*

10: Linear Regression and Correlation - *CC BY 4.0*
- 10.1: Introduction to Linear Regression and Correlation - *CC BY 4.0*
  - 10.1.1: Linear Equations - *CC BY 4.0*
    - 10.1.1E: Linear Equations (Exercises) - *CC BY 4.0*
  - 10.1.2: Scatter Plots - *CC BY 4.0*
    - 10.1.2E: Scatter Plots (Exercises) - *CC BY 4.0*
- 10.2: The Regression Equation and Correlation Coefficient - *CC BY 4.0*
  - 10.2E: The Regression Equation (Exercise) - *CC BY 4.0*
- 10.3: Testing for Significance Linear Correlation - *CC BY 4.0*
  - 10.3E: Testing the Significance of the Correlation Coefficient (Exercises) - *CC BY 4.0*
- 10.4: Prediction - *CC BY 4.0*
  - 10.4E: Prediction (Exercises) - *CC BY 4.0*
- 10.E: Linear Regression and Correlation (Exercises) - *CC BY 4.0*

11: Chi-Square Tests - *Undeclared*
- 11.1: Chi-Square Tests for Independence - *CC BY-NC-SA 3.0*
- 11.2: Chi-Square One-Sample Goodness-of-Fit Tests - *CC BY-NC-SA 3.0*

12: Analysis of Variance - *Undeclared*
- 12.1: F-Tests - *CC BY-NC-SA 3.0*
- 12.2: F-Tests in One-Way ANOVA - *CC BY-NC-SA 3.0*

Back Matter - *Undeclared*
- Index - *Undeclared*
- Glossary - *Undeclared*
- Detailed Licensing - *Undeclared*
- Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: Math 11: Elementary Statistics

Webpages: 85

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 3.0: 47.1% (40 pages)
- CC BY 4.0: 32.9% (28 pages)
- Undeclared: 18.8% (16 pages)
- CC BY-NC-SA 4.0: 1.2% (1 page)

By Page

- Math 11: Elementary Statistics - *CC BY-NC-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Statistics - *CC BY-NC-SA 3.0*
    - 1.1: Definitions of Statistics and Key Terms - *CC BY 4.0*
    - 1.2: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 1.3: Frequency, Frequency Tables, and Levels of Measurement - *CC BY 4.0*
      - 1.3.1: Experimental Design and Ethics - *CC BY 4.0*
  - 2: Data Displays - *Undeclared*
    - 2.1: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 2.2: Histogram - *CC BY-NC-SA 3.0*
    - 2.3: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs - *CC BY 4.0*
  - 3: Descriptive Statistics - *CC BY-NC-SA 3.0*
    - 3.1: Measures of Center - *CC BY-NC-SA 3.0*
    - 3.2: Measures of Variability - *CC BY-NC-SA 3.0*
    - 3.3: Relative Position of Data - *CC BY-NC-SA 3.0*
    - 3.4: The Empirical Rule and Chebyshev's Theorem - *CC BY-NC-SA 3.0*
  - 4: Probability Topics - *CC BY 4.0*
    - 4.1: Introduction - *CC BY 4.0*
      - 4.1.1: Terminology - *CC BY 4.0*
      - 4.1.2: Independent and Mutually Exclusive Events - *CC BY 4.0*
    - 4.2: Addition and Multiplication Rule of Probability - *CC BY 4.0*
    - 4.3: Conditional Probability using Contingency Tables - *CC BY 4.0*
    - 4.E: Probability Topics (Exericses) - *CC BY 4.0*
  - 5: Discrete Random Variables - *CC BY-NC-SA 3.0*
    - 5.1: Random Variables - *CC BY-NC-SA 3.0*
      - 5.1.1: Probability Distributions for Discrete Random Variables - *CC BY-NC-SA 3.0*
    - 5.2: The Binomial Distribution - *CC BY-NC-SA 3.0*
    - 5.E: Discrete Random Variables (Exercises) - *CC BY-NC-SA 3.0*
  - 6: Continuous Random Variables - *CC BY-NC-SA 3.0*
    - 6.1: The Standard Normal Distribution - *Undeclared*
      - 6.1.1: Continuous Random Variables - *CC BY-NC-SA 3.0*
      - 6.1.2: The Standard Normal Distribution - *CC BY-NC-SA 3.0*
    - 6.2: The General Normal Distribution - *CC BY-NC-SA 3.0*
    - 6.3: The Central Limit Theorem for Sample Means - *CC BY 4.0*
      - 6.3E: The Central Limit Theorem for Sample Means (Exercises) - *CC BY 4.0*
  - 7: Estimation - *CC BY-NC-SA 3.0*
    - 7.1: Estimation of a Population Proportion - *CC BY-NC-SA 3.0*
    - 7.2: Estimation of a Population Mean - *Undeclared*
      - 7.2.1: Large Sample Estimation of a Population Mean - *CC BY-NC-SA 3.0*
      - 7.2.2: Small Sample Estimation of a Population Mean - *CC BY-NC-SA 3.0*

- 7.3: Sample Size Considerations - *CC BY-NC-SA 3.0*
- 7.E: Estimation (Exercises) - *CC BY-NC-SA 3.0*

8: Testing Hypotheses - *CC BY-NC-SA 3.0*
- 8.1: The Elements of Hypothesis Testing - *CC BY-NC-SA 3.0*
- 8.2: Tests for a Population Mean - *CC BY-NC-SA 3.0*
  - 8.2.1: The Observed Significance of a Test - *CC BY-NC-SA 3.0*
  - 8.2.2: Small Sample Tests for a Population Mean - *CC BY-NC-SA 3.0*
- 8.3: Tests for a Population Proportion - *CC BY-NC-SA 3.0*
- 8.E: Testing Hypotheses (Exercises) - *CC BY-NC-SA 3.0*

9: Two-Sample Problems - *CC BY-NC-SA 3.0*
- 9.1: Two Population Proportions - *CC BY-NC-SA 3.0*
- 9.2: Two Population Means - Independent Samples - *Undeclared*
  - 9.2.1: Large, Independent Samples - *CC BY-NC-SA 3.0*
  - 9.2.2: Small, Independent Samples - *CC BY-NC-SA 3.0*
- 9.3: Two Population Means - Paired Samples - *CC BY-NC-SA 3.0*
- 9.4: Sample Size Considerations - *CC BY-NC-SA 3.0*
- 9.E: Two-Sample Problems (Exercises) - *CC BY-NC-SA 3.0*

10: Linear Regression and Correlation - *CC BY 4.0*
- 10.1: Introduction to Linear Regression and Correlation - *CC BY 4.0*
  - 10.1.1: Linear Equations - *CC BY 4.0*
    - 10.1.1E: Linear Equations (Exercises) - *CC BY 4.0*
  - 10.1.2: Scatter Plots - *CC BY 4.0*
    - 10.1.2E: Scatter Plots (Exercises) - *CC BY 4.0*
- 10.2: The Regression Equation and Correlation Coefficient - *CC BY 4.0*
  - 10.2E: The Regression Equation (Exercise) - *CC BY 4.0*
- 10.3: Testing for Significance Linear Correlation - *CC BY 4.0*
  - 10.3E: Testing the Significance of the Correlation Coefficient (Exercises) - *CC BY 4.0*
- 10.4: Prediction - *CC BY 4.0*
  - 10.4E: Prediction (Exercises) - *CC BY 4.0*
- 10.E: Linear Regression and Correlation (Exercises) - *CC BY 4.0*

11: Chi-Square Tests - *Undeclared*
- 11.1: Chi-Square Tests for Independence - *CC BY-NC-SA 3.0*
- 11.2: Chi-Square One-Sample Goodness-of-Fit Tests - *CC BY-NC-SA 3.0*

12: Analysis of Variance - *Undeclared*
- 12.1: F-Tests - *CC BY-NC-SA 3.0*
- 12.2: F-Tests in One-Way ANOVA - *CC BY-NC-SA 3.0*

Back Matter - *Undeclared*
- Index - *Undeclared*
- Glossary - *Undeclared*
- Detailed Licensing - *Undeclared*
- Detailed Licensing - *Undeclared*