

MATH 10: ELEMENTARY STATISTICS



Cliff Hwang
Mission College

Licensing

A detailed breakdown of this resource's licensing can be found in [Back Matter/Detailed Licensing](#).

Mission College

Math 10: Elementary Statistics

Cliff Hwang

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 adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

This text was compiled on 06/12/2025

TABLE OF CONTENTS

Licensing

1: Sampling and Data

- 1.0: Introduction
- 1.1: Definitions of Statistics, Probability, and Key Terms
- 1.2: Data, Sampling, and Variation in Data and Sampling
- 1.3: Frequency and Frequency Tables
- 1.4: Experimental Design and Ethics

2: Descriptive Statistics

- 2.0: Prelude to Descriptive Statistics
- 2.1: Line Graphs and Bar Graphs
- 2.2: Histograms and Time Series Graphs
- 2.3: Measures of the Location of the Data
- 2.4: Box Plots
- 2.5: Measures of the Center of the Data
- 2.6: Skewness and the Mean, Median, and Mode
- 2.7: Measures of the Spread of the Data

3: Linear Regression and Correlation

- 3.0: Prelude to Linear Regression and Correlation
- 3.1: Linear Equations
- 3.2: Scatter Plots
- 3.3: The Regression Equation
- 3.4: Prediction

4: Probability Topics

- 4.0: Introduction
- 4.1: Terminology

5: Discrete Random Variables

- 5.0: Prelude to Discrete Random Variables
- 5.1: Probability Distribution Function (PDF) for a Discrete Random Variable
- 5.2: Mean and Standard Deviation
- 5.3: Binomial Distribution

6: Continuous Random Variables

- 6.0: Introduction
- 6.1: Continuous Probability Functions
- 6.2: The Uniform Distribution

7: The Normal Distribution

- 7.0: Prelude to The Normal Distribution
- 7.1: The Standard Normal Distribution

- [7.2: Using the Normal Distribution](#)

8: The Central Limit Theorem

- [8.0: Prelude to the Central Limit Theorem](#)
- [8.1: The Central Limit Theorem for Sample Means \(Averages\)](#)
- [8.2: The Central Limit Theorem for Sums](#)
- [8.3: Using the Central Limit Theorem](#)

9: Confidence Intervals

- [9.0: Prelude to Confidence Intervals](#)
- [9.1: A Single Population Mean using the Normal Distribution](#)
- [9.2: A Single Population Mean using the Student t-Distribution](#)
- [9.3: A Population Proportion](#)
- [9.S: Confidence Intervals \(Summary\)](#)

10: Hypothesis Testing with One Sample

- [10.0: Prelude to Hypothesis Testing](#)
- [10.1: Null and Alternative Hypotheses](#)
- [10.2: Outcomes and the Type I and Type II Errors](#)
- [10.3: Distribution Needed for Hypothesis Testing](#)
- [10.4: Rare Events, the Sample, Decision and Conclusion](#)
- [10.5: Additional Information and Full Hypothesis Test Examples](#)

11: Hypothesis Testing with Two Samples

- [11.1: Comparing Two Independent Population Means \(Hypothesis test\)](#)
- [11.2: Comparing Two Independent Population Proportions \(Hypothesis test\)](#)
- [11.3: Matched or Paired Samples](#)

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

Licensing

A detailed breakdown of this resource's licensing can be found in [Back Matter/Detailed Licensing](#).

CHAPTER OVERVIEW

1: Sampling and Data

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

[1.0: Introduction](#)

[1.1: Definitions of Statistics, Probability, and Key Terms](#)

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

[1.3: Frequency and Frequency Tables](#)

[1.4: Experimental Design and Ethics](#)

Contributors and Attributions

Template:ContribOpenStax

This page titled [1: Sampling and Data](#) 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.

1.0: Introduction

Learning Objectives

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

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

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



Figure 1.0.1 We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

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

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

This page titled [1.0: 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.

1.1: Definitions of Statistics, Probability, and Key Terms

The science of statistics deals with the collection, analysis, interpretation, and presentation of data. We see and use data in our everyday lives.

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; 6.5; 6.5; 6.5; 6.5; 7; 7; 8; 8; 9

The dot plot for this data would be as follows:

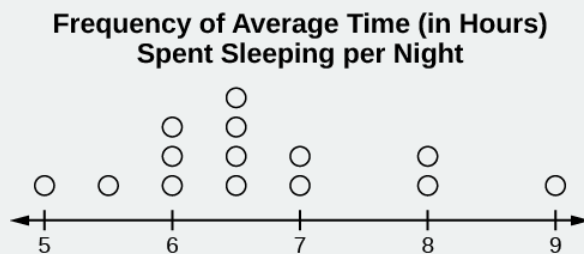


Figure 1.1.1

- 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.

Probability

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

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or

not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

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

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. 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 and ask them how much money they spent on school supplies that do not include books.

Answer

- The **population** is all first year students attending ABC College this term.
- The **sample** is the 100 first year students that were surveyed.
- The **parameter** is the average (mean) amount of money spent (excluding books) by all first year college students at ABC College this term.

- The **statistic** is the average (mean) amount of money spent (excluding books) by the 100 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 100 first year students surveyed. For example, three of those students could have spent \$150, \$200, and \$225, respectively

? 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.

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.

? Exercise 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

- all students who attended the college last year
- the cumulative GPA of one student who graduated from the college last year
- 3.65, 2.80, 1.50, 3.90
- a group of students who graduated from the college last year, randomly selected
- the average cumulative GPA of students who graduated from the college last year
- all students who graduated from the college last year
- 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.

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

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

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, Probability, 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?

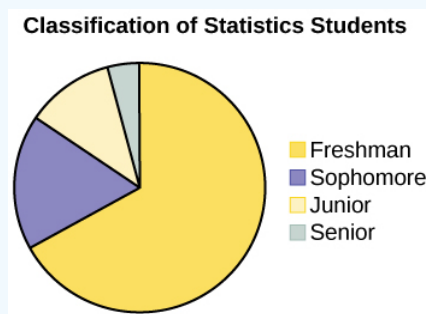


Figure 1.2.1

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.

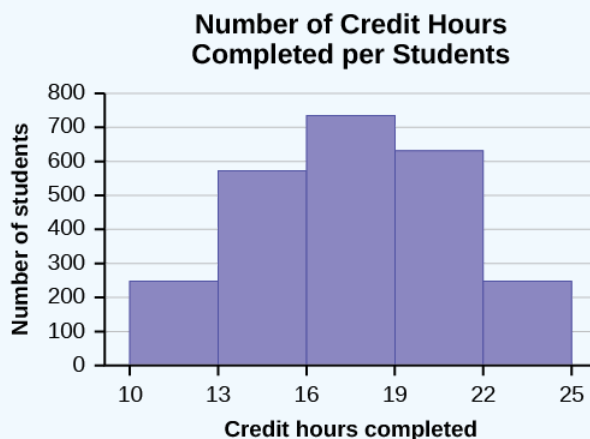


Figure 1.2.2

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.

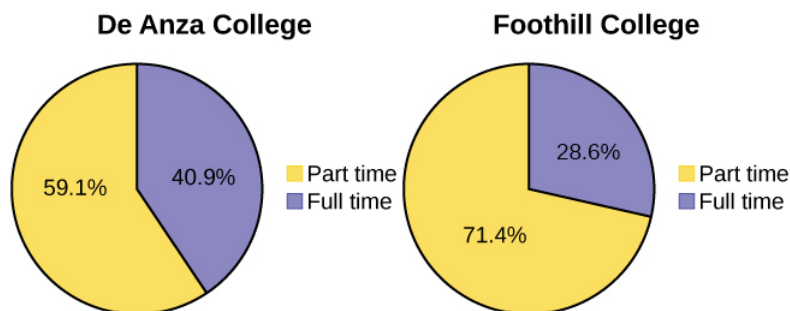


Figure 1.2.3: Pie Charts

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.

Student Status

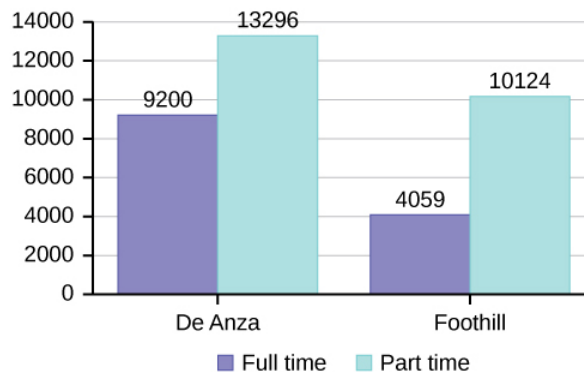


Figure 1.2.4: Bar chart

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%

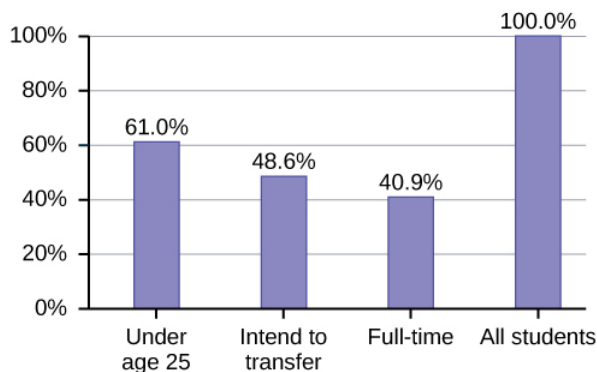


Figure 1.2.5: Bar chart of data in Table 1.2.2.

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

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%

	Frequency	Percent
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%

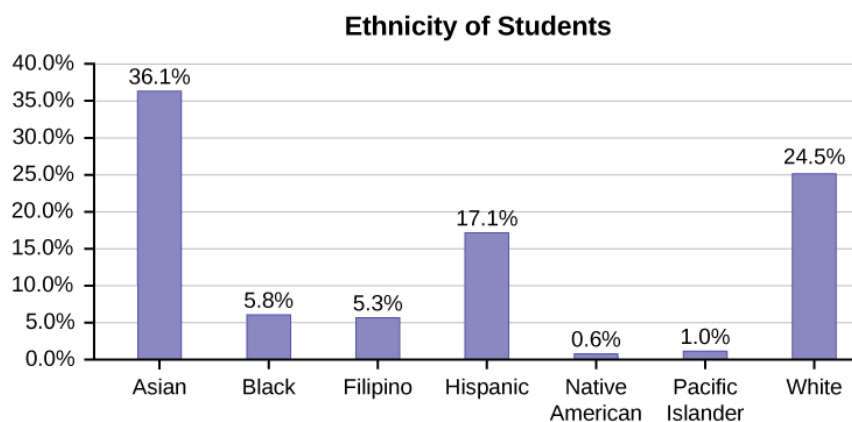


Figure 1.2.6: Enrollment of De Anza College (Spring 2010)

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 1.2.7 can be difficult to understand visually. The graph in Figure 1.2.8 is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

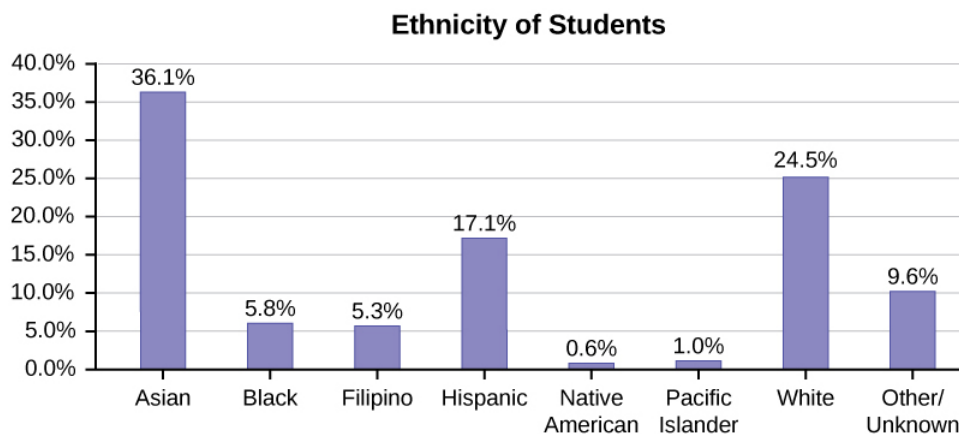


Figure 1.2.7: Bar Graph with Other/Unknown Category

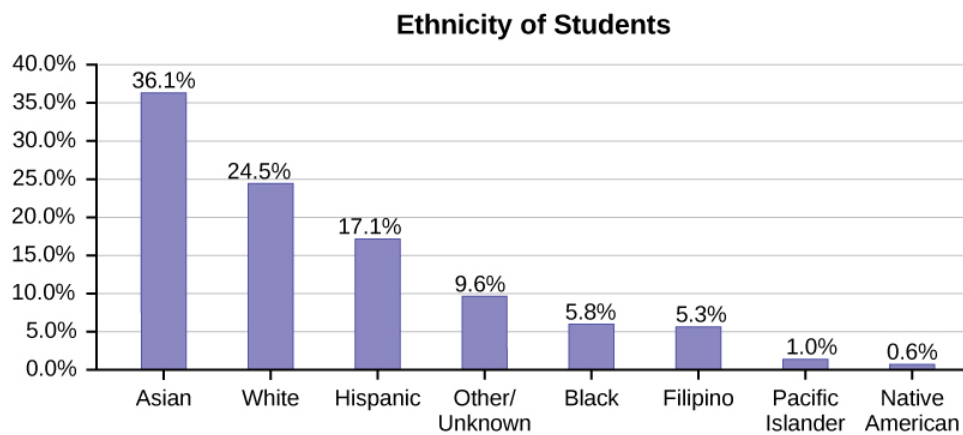


Figure 1.2.8: Pareto Chart With Bars Sorted by Size

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 1.2.9 is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 1.2.9.

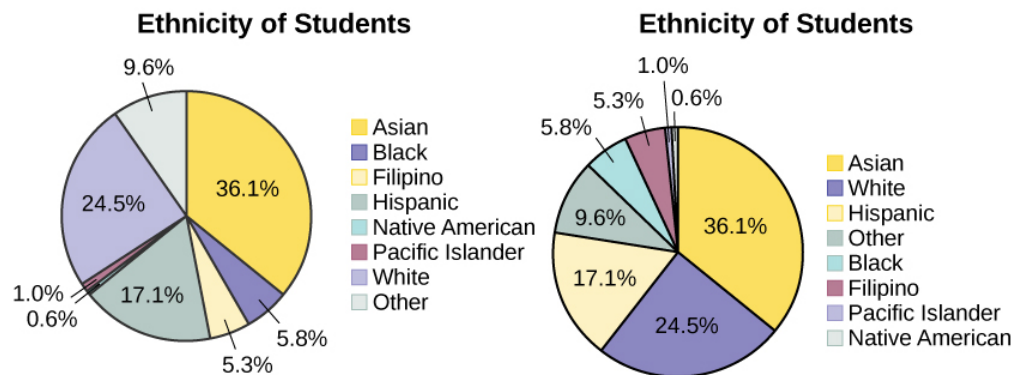


Figure 1.2.9.

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

ID	Name	ID	Name	ID	Name
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cunningham	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 Cunningham, 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.

```
randInt(0,30) 29
randInt(0,30) 28
randInt(0,30) 4
```

Figure 1.2.10

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 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.
- 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.
- 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.
- 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.
- 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.

- 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.

- o 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.
 - o Repeat for columns two through six.
 - o 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.
- o Press MATH and arrow over to PRB.
 - o Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
 - o The two numbers are for two of the columns.
 - o The quiz scores (20 of them) in these 2 columns are the cluster sample.
- c. Create a simple random sample of 15 quiz scores.
- o Use the numbering one through 60.
 - o Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
 - o Press ENTER 15 times and record the numbers.
 - o Record the quiz scores that correspond to these numbers.
 - o These 15 quiz scores are the systematic sample.
- d. Create a systematic sample of 12 quiz scores.
- o Use the numbering one through 60.
 - o Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
 - o 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 1.2.2

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

- To find the average GPA of all students in a university, use all honor students at the university as the sample.
- 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.
- 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.
- 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.
- 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 1.2.3

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/Relationships...the-President
5. Dominic Lusinch, “‘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 $k = (\text{number of individuals in the population}) / (\text{number of individuals needed in the sample})$. Choose every k th 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 and Frequency Tables

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.

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 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

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
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. Because of rounding, the relative frequency column may not always sum to one. However, it 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
59.95–61.95	5	$\frac{5}{100} = 0.05$
61.95–63.95	3	$\frac{3}{100} = 0.03$
63.95–65.95	15	$\frac{15}{100} = 0.15$
65.95–67.95	40	$\frac{40}{100} = 0.40$
67.95–69.95	17	$\frac{17}{100} = 0.17$
69.95–71.95	12	$\frac{12}{100} = 0.12$
71.95–73.95	7	$\frac{7}{100} = 0.07$
73.95–75.95	1	$\frac{1}{100} = 0.01$
	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

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

Answer

- 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.
- 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.

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

Table 1.3.5

Rainfall (Inches)	Frequency	Relative Frequency
2.95–4.97	6	$\frac{6}{50} = 0.12$
4.97–6.99	7	$\frac{7}{50} = 0.14$
6.99–9.01	15	$\frac{15}{50} = 0.30$
9.01–11.03	8	$\frac{8}{50} = 0.16$
11.03–13.05	9	$\frac{9}{50} = 0.18$
13.05–15.07	5	$\frac{5}{50} = 0.10$
	Total = 50	Total = 1.00

Answer

- 0.56 or 56
- $0.30 + 0.16 + 0.18 = 0.64$ 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.

- The percentage of heights that are from 67.95 to 71.95 inches is: ____.
- The percentage of heights that are from 67.95 to 73.95 inches is: ____.
- The percentage of heights that are more than 65.95 inches is: ____.
- The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
- What kind of data are the heights?
- 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

- 29%
- 36%
- 77%
- 87
- quantitative continuous
- 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

$$6 + 7 + 15 = 28 \text{ towns}$$

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:

- What percentage of the students in your class have no siblings?
- What percentage of the students have from one to three siblings?
- What percentage of the students have fewer than three siblings?

✓ Exercise 1.3.5

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
3	3	$\frac{3}{19}$
4	1	$\frac{1}{19}$
5	3	$\frac{3}{19}$
7	2	$\frac{2}{19}$
10	3	$\frac{3}{19}$
12	2	$\frac{2}{19}$
13	1	$\frac{1}{19}$
15	1	$\frac{1}{19}$
18	1	$\frac{1}{19}$
20	1	$\frac{1}{19}$

- Is the table correct? If it is not correct, what is wrong?
- 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.
- What fraction of the people surveyed commute five or seven miles?
- 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

- No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
- False. The frequency for three miles should be one; for two miles (left out), two.
- $\frac{5}{19}$
- $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

? Exercise 1.3.6

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 1.3.7

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.

- What is the frequency of deaths measured from 2006 through 2009?
- What percentage of deaths occurred after 2009?
- What is the relative frequency of deaths that occurred in 2003 or earlier?
- What is the percentage of deaths that occurred in 2004?
- What kind of data are the numbers of deaths?
- 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

- 97,118 (11.8%)
- 41.6%
- $67,092/823,356$ or 0.081 or 8.1 %
- 27.8%
- Quantitative discrete
- Quantitative continuous

? Exercise 1.3.8

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
------	-------------------------	------	-------------------------

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.

- What is the frequency of deaths measured from 2000 through 2004?
- What percentage of deaths occurred after 2006?
- What is the relative frequency of deaths that occurred in 2000 or before?
- What is the percentage of deaths that occurred in 2011?
- What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Answer

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

References

- “State & County QuickFacts,” U.S. Census Bureau. quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).
- “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).
- “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).

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.

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.

Glossary

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 and Frequency 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.

1.4: 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 1.4.1

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 *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 1.4.2

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.

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

Answer

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

✓ Example 1.4.3

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 1.4.4

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?

- Describe the explanatory and response variables in the study.
- What are the treatments?
- What should you consider when selecting participants?
- 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?
- Identify any lurking variables that could interfere with this study.
- How can blinding be used in this study?

Answer

- Explanatory: presence of distraction from texting; response: response time measured in seconds
- Driving without distraction and driving while texting
- 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 1.4.5

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 1.4.6

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/nutrition-source/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), <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-aspiring-study-300443> (accessed May 1, 2013).
4. The Data and Story Library, lib.stat.cmu.edu/DASL/Stories/StoryLearning.html (accessed May 1, 2013).
5. M.L. Jakszon 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/earthquakes/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-consumer-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 1.4.7

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 1.4.7

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

³ "Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," *Tilburg 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 [1.4: 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: Descriptive Statistics

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

[2.0: Prelude to Descriptive Statistics](#)

[2.1: Line Graphs and Bar Graphs](#)

[2.2: Histograms and Time Series Graphs](#)

[2.3: Measures of the Location of the Data](#)

[2.4: Box Plots](#)

[2.5: Measures of the Center of the Data](#)

[2.6: Skewness and the Mean, Median, and Mode](#)

[2.7: Measures of the Spread of the Data](#)

This page titled [2: Descriptive Statistics](#) 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.

2.0: Prelude to Descriptive Statistics

Learning Objectives

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

- Display data graphically and interpret graphs: histograms and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.



Figure 2.0.1: When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

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

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

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

This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The Texas Instruments (TI) website provides additional instructions for using these calculators.

This page titled [2.0: Prelude to Descriptive Statistics](#) 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.

2.1: Line Graphs and Bar Graphs

A 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 2.1.7

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 the table below. Construct a corresponding line graph.

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

Answer

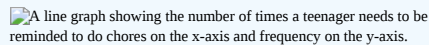
A line graph showing the number of times a teenager needs to be reminded to do chores on the x-axis and frequency on the y-axis.

Figure 2.1.1: 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 2.1.8

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 the table below. Construct a line graph.

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

Answer

Figure 2.1.2: 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 2.1.9 has age groups represented on the **x-axis** and proportions on the **y-axis**.

? Example 2.1.9

By the end of 2011, Facebook had over 146 million users in the United States. The table below 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 2.1.3: 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 2.1.10

The population in Park City is made up of children, working-age adults, and retirees. The table below shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Answer

Figure 2.1.4: 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 2.1.11

The columns in the table below contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examinee 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%

Answer

Figure 2.1.5: 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 2.1.12

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 2.1.6: 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 **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/facebook-demographics-revisited-2001-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 [2.1: 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.

2.2: Histograms and Time Series Graphs

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

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

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

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

then:

$$RF = \frac{f}{n} \quad (2.2.1)$$

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

To construct a histogram, first decide how many bars or intervals, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 ($6.1 - 0.05 = 6.05$). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 ($1.5 - 0.005 = 1.495$). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 ($1.0 - 0.0005 = 0.9995$). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 ($2 - 0.5 = 1.5$). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

Example 2.2.1

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

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71

72; 72; 72.5; 72.5; 73; 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

$60 - 0.05 = 59.95$ which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

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

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

$$\frac{74.05 - 59.95}{8} = 1.76 \quad (2.2.2)$$

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

The boundaries are:

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

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

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

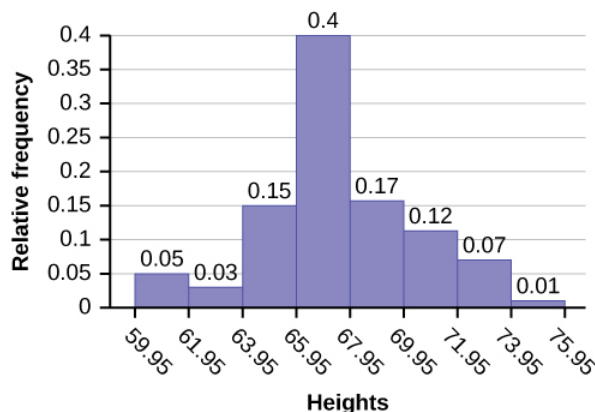


Figure 2.2.1: Histogram of something

Exercise 2.2.2

The following data are the shoe sizes of 50 male students. The sizes are discrete data since shoe size is measured in whole and half units only. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.

9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5
 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Answer

Smallest value: 9

Largest value: 14

Convenient starting value: $9 - 0.05 = 8.95$

Convenient ending value: $14 + 0.05 = 14.05$

$$\frac{14.05 - 8.95}{6} = 0.85$$

The calculations suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Example 2.2.3

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1

2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3

4; 4; 4; 4; 4

5; 5; 5; 5

6; 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____.

Answer

Calculate the number of bars as follows:

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

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the x-axis and the frequency on the y-axis.

Figure 2.2.2: Histogram consists of 6 bars with the y-axis in increments of 2 from 0-16 and the x-axis in intervals of 1 from 0.5-6.5.

Note

Go to [\[link\]](#). There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for [Example](#).

- Press Y=. Press CLEAR to delete any equations.

1.5
1.5 to 2.5
2.5 to 3.5

9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Figure 2.2.3: This is a histogram that matches the supplied data. The x-axis consists of 5 bars in intervals of 5 from 0 to 25. The y-axis is marked in increments of 1 from 0 to 10. The x-axis shows the number of hours spent playing video games on the weekends, and the y-axis shows the number of students.

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Exercise 2.2.6

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22; 35; 15; 26; 40; 28; 18; 20; 25; 34; 39; 42; 24; 22; 19; 27; 22; 34; 40; 20; 38 and 28

Use 10–19 as the first interval.

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals.

Constructing a Time Series Graph

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

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

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

Example 2.2.7

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

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0

Year	Aug	Sep	Oct	Nov	Dec	Annual
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Answer

Figure 2.2.7: This is a times series graph that matches the supplied data. The x-axis shows years from 2003 to 2012, and the y-axis shows the annual CPI.

Exercise 2.2.8

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO₂ emissions for the United States.

CO ₂ Emissions			
	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Answer

Figure 2.2.8: This is a times series graph that matches the supplied data. The x-axis shows years from 2003 to 2012, and the y-axis shows the annual CPI.

Uses of a Time Series Graph

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

Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time. Glossary

References

1. Data on annual homicides in Detroit, 1961–73, from Gunst & Mason’s book ‘Regression Analysis and its Application’, Marcel Dekker
2. “Timeline: Guide to the U.S. Presidents: Information on every president’s birthplace, political party, term of office, and more.” Scholastic, 2013. Available online at www.scholastic.com/teachers/a...-us-presidents (accessed April 3, 2013).
3. “Presidents.” Fact Monster. Pearson Education, 2007. Available online at <http://www.factmonster.com/ipka/A0194030.html> (accessed April 3, 2013).
4. “Food Security Statistics.” Food and Agriculture Organization of the United Nations. Available online at <http://www.fao.org/economic/ess/ess-fs/en/> (accessed April 3, 2013).
5. “Consumer Price Index.” United States Department of Labor: Bureau of Labor Statistics. Available online at <http://data.bls.gov/pdq/SurveyOutputServlet> (accessed April 3, 2013).
6. “CO2 emissions (kt).” The World Bank, 2013. Available online at <http://databank.worldbank.org/data/home.aspx> (accessed April 3, 2013).
7. “Births Time Series Data.” General Register Office For Scotland, 2013. Available online at www.gro-scotland.gov.uk/statistics/time-series.html (accessed April 3, 2013).
8. “Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en> (accessed April 3, 2013).
9. Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.
10. “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).

Frequency

the number of times a value of the data occurs

Histogram

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

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

This page titled [2.2: Histograms and Time Series 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.

2.3: Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**. Quartiles are special percentiles. The first quartile, Q_1 , is the same as the 25th percentile, and the third quartile, Q_3 , is the same as the 75th percentile. The median, M , is called both the second quartile and the 50th percentile.

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

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

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

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

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

Ordered from smallest to largest:

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

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7 \quad (2.3.1)$$

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

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1 , is the middle value of the lower half of the data, and the third quartile, Q_3 , is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

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

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

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

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

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

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

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

$$IQR = Q_3 - Q_1 \quad (2.4.1)$$

The *IQR* can help to determine potential **outliers**. A value is suspected to be a potential outlier if it is less than $(1.5)(IQR)$ below the first quartile or more than $(1.5)(IQR)$ above the third quartile. Potential outliers always require further investigation.

Definition: Outliers

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

✓ Exercise 2.3.1

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

Answer

Order the data from smallest to largest.

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

$$M = 488,800$$

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

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

$$IQR = 649,000 - 308,750 = 340,250$$

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

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

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

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

? Exercise 2.3.2

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.
\$33,000; \$64,500; \$28,000; \$54,000; \$72,000; \$68,500; \$69,000; \$42,000; \$54,000; \$120,000; \$40,500

Answer

Order the data from smallest to largest.

\$28,000; \$33,000; \$40,500; \$42,000; \$54,000; \$54,000; \$64,500; \$68,500; \$69,000; \$72,000; \$120,000

Median = \$54,000

$$Q_1 = \$40,500$$

$$Q_3 = \$69,000$$

$$IQR = \$69,000 - \$40,500 = \$28,500$$

$$(1.5)(IQR) = (1.5)(\$28,500) = \$42,750$$

$$Q_1 - (1.5)(IQR) = \$40,500 - \$42,750 = -\$2,250$$

$$Q_3 + (1.5)(IQR) = \$69,000 + \$42,750 = \$111,750$$

No salary is less than $-\$2,250$. However, \$120,000 is more than \$111,750, so \$120,000 is a potential outlier.

✓ Exercise 2.3.3

For the two data sets in the [test scores example](#), find the following:

- The interquartile range. Compare the two interquartile ranges.
- Any outliers in either set.

Answer

The five number summary for the day and night classes is

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

- The IQR for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

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

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

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

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

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

Night class outliers are calculated as:

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

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

? Exercise 2.3.4

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class A

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Test Scores for Class B

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Answer

Class A

Order the data from smallest to largest.

65; 66; 67; 69; 69; 76; 77; 77; 79; 80; 81; 83; 85; 89; 90; 91; 94; 96; 98; 99

$$\text{Median} = \frac{80 + 81}{2} = 80.5$$

$$Q_1 = \frac{69 + 76}{2} = 72.5$$

$$Q_3 = \frac{90 + 91}{2} = 90.5$$

$$IQR = 90.5 - 72.5 = 18$$

Class B

Order the data from smallest to largest.

68; 68; 70; 71; 72; 73; 75; 78; 79; 80; 80; 90; 90; 92; 92; 95; 95; 97; 99; 100

$$Median = \frac{80 + 80}{2} = 80$$

$$Q_1 = \frac{72 + 73}{2} = 72.5$$

$$Q_3 = \frac{92 + 95}{2} = 93.5$$

$$IQR = 93.5 - 72.5 = 21$$

The data for Class B has a larger *IQR*, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class B are more spread out and not clustered about the median.

✓ Exercise 2.3.5

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

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

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

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

? Exercise 2.3.6

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Answer

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

✓ Exercise 2.3.7

Using the table above in Example 2.3.3

- Find the 80th percentile.
- Find the 90th percentile.
- Find the first quartile. What is another name for the first quartile?

Answer

Using the data from the frequency table, we have:

- The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th and 41st values. The 80th percentile = $\frac{8+9}{2} = 8.5$
- The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
- Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

? Exercise 2.3.8

Refer to the table above in Exercise 2.3.3. Find the third quartile. What is another name for the third quartile?

Answer

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

📌 COLLABORATIVE STATISTICS

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

- How many students were surveyed?
- What kind of sampling did you do?
- Construct two different histograms. For each, starting value = _____ ending value = _____.
- Find the median, first quartile, and third quartile.

- e. Construct a table of the data to find the following:
- the 10th percentile
 - the 70th percentile
 - the percent of students who own less than four sweaters

A Formula for Finding the k th Percentile

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

- k = the k th percentile. It may or may not be part of the data.
- i = the index (ranking or position of a data value)
- n = the total number of data

Order the data from smallest to largest.

$$\text{Calculate } i = \frac{k}{100}(n + 1)$$

If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.

If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

✓ Exercise 2.3.9

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

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

- Find the 70th percentile.
- Find the 83rd percentile.

Answer

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

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

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

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

? Exercise 2.3.10

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

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

Calculate the 20th percentile and the 55th percentile.

Answer

$k = 20$. Index $= i = \frac{k}{100}(n + 1) = \frac{20}{100}(29 + 1) = 6$. The age in the sixth position is 27. The 20th percentile is 27 years.

$k = 55$. Index $= i = \frac{k}{100}(n + 1) = \frac{55}{100}(29 + 1) = 16.5$. Round down to 16 and up to 17. The age in the 16th position is 52 and the age in the 17th position is 55. The average of 52 and 55 is 53.5. The 55th percentile is 53.5 years.

Note 2.4.2

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

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

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x + 0.5y}{n}(100)$. Then round to the nearest integer.

✓ Exercise 2.3.11

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

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

- Find the percentile for 58.
- Find the percentile for 25.

Answer

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

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

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

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

? Exercise 2.3.12

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

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

Find the percentiles for 47 and 31.

Answer

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

$$x = 15 \text{ and } y = 1. \frac{x + 0.5y}{n}(100) = \frac{15 + 0.5(1)}{30}(100) = 51.67. 47 \text{ is the } 52^{\text{nd}} \text{ percentile.}$$

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

$$x = 8 \text{ and } y = 2. \frac{x + 0.5y}{n}(100) = \frac{8 + 0.5(2)}{30}(100) = 30. 31 \text{ is the } 30^{\text{th}} \text{ percentile.}$$

Interpreting Percentiles, Quartiles, and Median

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

15th percentile.

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

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

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

GUIDELINE

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

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

? Exercise 2.3.13

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

Answer

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

? Exercise 2.3.14

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Answer

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

? Exercise 2.3.15

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Answer

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

? Exercise 2.3.16

On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Answer

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

✓ Exercise 2.3.17

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

Answer

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

? Exercise 2.3.18

During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Answer

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

✓ Exercise 2.3.19

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

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

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

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

Determine the following five values. (this is also called a 5 number summary)

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

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

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

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120 \quad (2.3.2)$$

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

- Min = 0
- $Q_1 = 20$
- Med = 35
- $Q_3 = 60$
- Max = 120

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

References

1. Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at usatoday30.usatoday.com/news/...sus/55029100/1 (accessed April 3, 2013).
2. Data from the United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/> (accessed April 3, 2013).
3. "1990 Census." United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/main/www/cen1990.html> (accessed April 3, 2013).
4. Data from *San Jose Mercury News*.
5. Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Review

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

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

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

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

- k = the k^{th} percentile,
- n = total number of data.

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

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

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

n = total number of data

Glossary

Interquartile Range

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

Outlier

an observation that does not fit the rest of the data

Percentile

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

Quartiles

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

This page titled [2.3: Measures of the Location of the Data](#) 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.

- [2.4: Measures of the Location of the Data](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

2.4: Box Plots

Box plots (also called *box-and-whisker plots* or *box-whisker plots*) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

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

You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

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

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

See the calculator instructions on the TI web site or in the appendix.

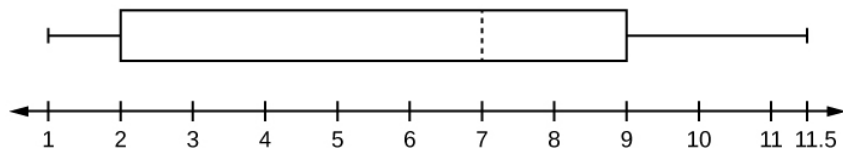


Figure 2.4.1

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

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

✓ Example 2.4.1

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

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

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

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

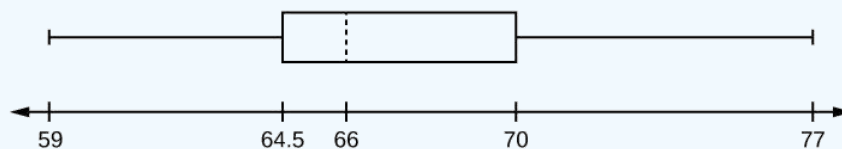


Figure 2.4.2

- a. Each quarter has approximately 25% of the data.
- b. The spreads of the four quarters are $64.5 - 59 = 5.5$ (first quarter), $66 - 64.5 = 1.5$ (second quarter), $70 - 66 = 4$ (third quarter), and $77 - 70 = 7$ (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
- c. Range = maximum value - the minimum value = $77 - 59 = 18$
- d. Interquartile Range: $IQR = Q_3 - Q_1 = 70 - 64.5 = 5.5$.
- e. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
- f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Calculator

To find the minimum, maximum, and quartiles:

Enter data into the list editor (Press STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1 : First quartile = 64.5.

Q_2 : Second quartile or median = 66.

Q_3 : Third quartile = 70.

To construct the box plot:

Press 4: Plotsoff. Press ENTER.

Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.

Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

Press TRACE, and use the arrow keys to examine the box plot.

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:

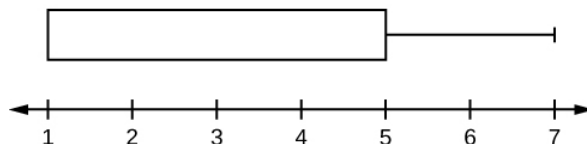


Figure 2.4.4

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

✓ Example 2.4.2

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

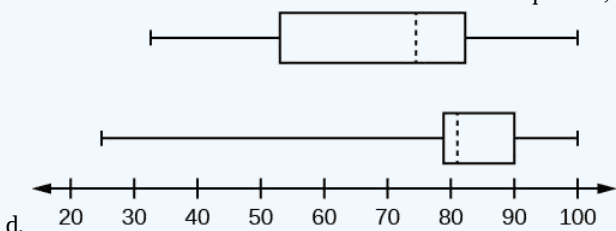
Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

- Find the smallest and largest values, the median, and the first and third quartile for the day class.
- Find the smallest and largest values, the median, and the first and third quartile for the night class.
- For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
- Create a box plot for each set of data. Use one number line for both box plots.
- Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Answer

- Min = 32
 - $Q_1 = 56$
 - $M = 74.5$
 - $Q_3 = 82.5$
 - Max = 99
- Min = 25.5
 - $Q_1 = 78$
 - $M = 81$
 - $Q_3 = 89$
 - Max = 98
- Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:



- The first data set has the wider spread for the middle 50% of the data. The *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

? Example 2.4.3

The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Answer

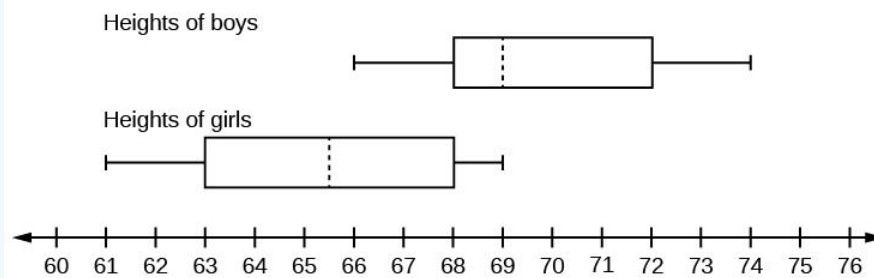


Figure 2.4.6

IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

✓ Example 2.4.4

Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1 : 15
- Med: 95
- Q_3 : 490
- Max: 790

The following graph shows the box-and-whisker plot.

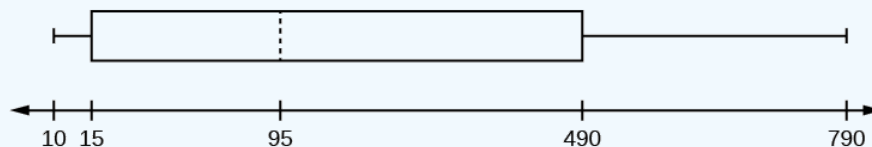


Figure 2.4.7

? Example 2.4.5

Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0; 5; 5; 15; 30; 30; 45; 50; 50; 60; 75; 110; 140; 240; 330

Answer

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the boxplot are:

- Min: 0
- Q_1 : 15
- Med: 50
- Q_3 : 110
- Max: 330

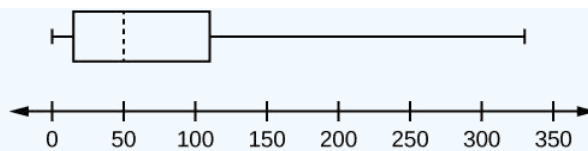


Figure 2.4.8

Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

? Example 2.4.7

Construct a box plot below. Use a ruler to measure and scale accurately.

? Example 2.4.8

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Answer

More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

Bringing It Together

? Example 2.4.9

Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Age Group	Percent of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

- Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
- What percentage of the community is under age 35?
- Which box plot most resembles the information above?

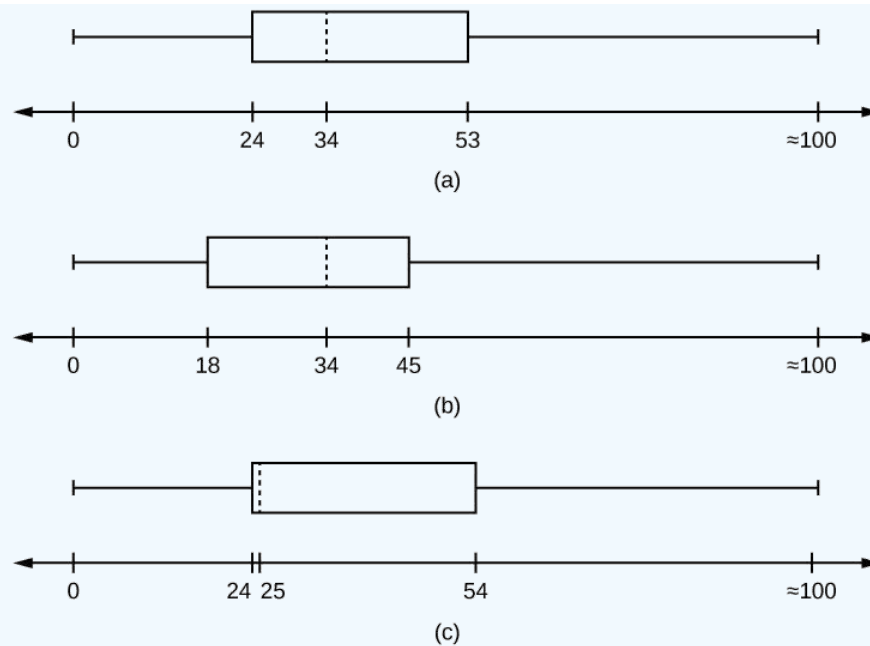


Figure 2.4.9.

Answer

- For graph, check student's solution.
- 49.7% of the community is under the age of 35.
- Based on the information in the table, graph (a) most closely represents the data.

Glossary

Box plot

a graph that gives a quick picture of the middle 50% of the data

First Quartile

the value that is the median of the of the lower half of the ordered data set

Frequency Polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Paired Data Set

two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

References

1. Data from *West Magazine*.

This page titled [2.4: Box 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.

2.5: Measures of the Center of the Data

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

Note

The ideal gas law is easy to remember and apply in solving problems, as long as you get the **proper values**. The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced " x bar"): $\bar{x}$.

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

To see that both ways of calculating the mean are the same, consider the sample:

1; 1; 1; 2; 2; 3; 4; 4; 4; 4

$$\bar{x} = \frac{1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4}{11} = 2.7 \quad (2.5.1)$$

$$\bar{x} = \frac{3(1) + 2(2) + 1(3) + 5(4)}{11} = 2.7 \quad (2.5.2)$$

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

You can quickly find the location of the median by using the expression

$$\frac{n+1}{2} \quad (2.5.3)$$

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \quad (2.5.4)$$

The median is the 49th value in the ordered data. If the total number of data values is 100, then

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \quad (2.5.5)$$

The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

✓ Example 2.5.1

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

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

Calculate the mean and the median.

Answer

The calculation for the mean is:

$$\bar{x} = \frac{[3 + 4 + (8)(2) + 10 + 11 + 12 + 13 + 14 + (15)(2) + (16)(2) + \dots + 35 + 37 + 40 + (44)(2) + 47]}{40} = 23.6 \quad (2.5.6)$$

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

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \quad (2.5.7)$$

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

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

$$M = \frac{24 + 24}{2} = 24 \quad (2.5.8)$$

Calculator

To find the mean and the median:

Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.

Enter data into the list editor. Press STAT 1:EDIT.

Put the data values into list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

$$\bar{x} = 23.6, M = 24$$

? Example 2.5.2

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 11; 12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21; 21; 22; 22; 23; 24; 24; 24; 24

Answer

Mean:

$$3 + 4 + 5 + 7 + 7 + 7 + 7 + 8 + 8 + 9 + 9 + 10 + 10 + 10 + 10 + 10 + 11 + 12 + 12 + 13 + 14 + 14 + 15 + 15 + 17 + 17 + 18 + 19 + 19 + 19 + 21 + 21 + 22 + 22 + 23 + 24 + 24 + 24 + 24 = 544$$

$$\frac{544}{39} = 13.95 \quad (2.5.9)$$

Median: Starting at the smallest value, the median is the 20th term, which is 13.

✓ Example 2.5.3

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

Answer

$$\bar{x} = \frac{5,000,000 + 49(30,000)}{50} = 129,400 \quad (2.5.10)$$

$$M = 30,000$$

(There are 49 people who earn \$30,000 and one person who earns \$5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

? Example 2.5.4

In a sample of 60 households, one house is worth \$2,500,000. Half of the rest are worth \$280,000, and all the others are worth \$315,000. Which is the better measure of the "center": the mean or the median?

Answer

The median is the better measure of the "center" than the mean because 59 of the values are \$280,000 and one is \$2,500,000. The \$2,500,000 is an outlier. Either \$280,000 or \$315,000 gives us a better sense of the middle of the data.

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

✓ Example 2.5.5

Statistics exam scores for 20 students are as follows:

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

Find the mode.

Answer

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

? Example 2.5.6

The number of books checked out from the library from 25 students are as follows:

0; 0; 0; 1; 2; 3; 3; 4; 4; 5; 5; 7; 7; 7; 8; 8; 9; 10; 10; 11; 11; 12; 12

Find the mode.

Answer

The most frequent number of books is 7, which occurs four times. Mode = 7.

✓ Example 2.5.7

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

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

The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

? Example 2.5.8

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is \$25,000 and occurs 150 times out of 301. The median is \$50,000 and the mean is \$47,500. What would be the best measure of the "center"?

Answer

Because \$25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don't represent what most people make at the factory.

📌 The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample is very likely to get closer and closer to μ . This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a sampling distribution as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A statistic is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ .

References

1. Data from The World Bank, available online at <http://www.worldbank.org> (accessed April 3, 2013).
2. "Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en> (accessed April 3, 2013).

Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occurring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$$\mu = \frac{\sum fm}{\sum f} \quad (2.5.11)$$

where f = interval frequencies and m = interval midpoints.

Glossary

Frequency Table

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

Mean

a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic

mean.' By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ)

is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median

a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint

the mean of an interval in a frequency table

Mode

the value that appears most frequently in a set of data

This page titled [2.5: Measures of the Center of the Data](#) 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.

- [2.6: Measures of the Center of the Data](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

2.6: Skewness and the Mean, Median, and Mode

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

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

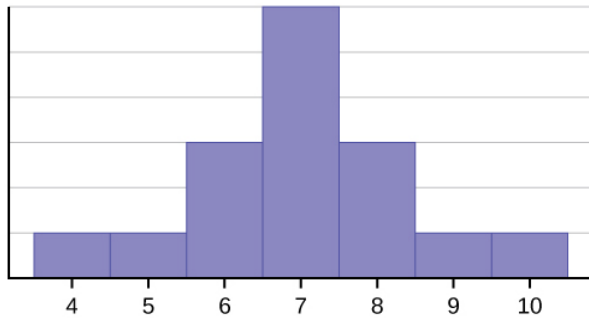


Figure 2.6.1

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

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

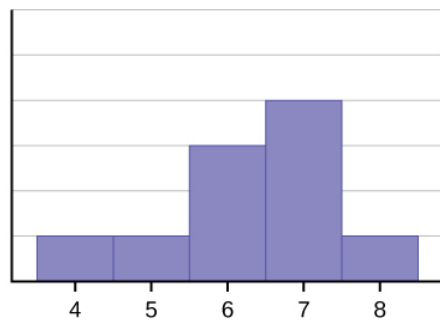


Figure 2.6.2

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

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

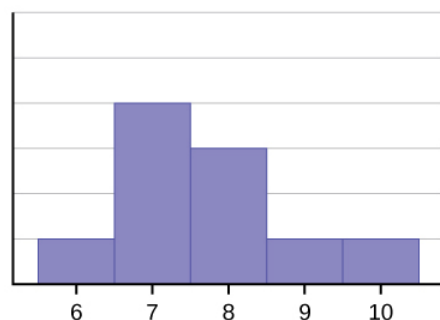


Figure 2.6.3

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

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

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

✓ Example 2.6.1

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

- Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2
- Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1
- Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

- Make a dot plot for the three authors and compare the shapes.
- Calculate the mean for each.
- Calculate the median for each.
- Describe any pattern you notice between the shape and the measures of center.

Solution

a. Figure 2.6.4 This dot plot matches the supplied data for Terry. The plot uses a number line from 1 to 10. It shows one x over 1, two x's over 2, four x's over 3, one x over 4, one x over 7, and one x over 9. There are no x's over the numbers 5, 6, 8, and 10.

Caption here. (Copyright; author via source)

Caption here. (Copyright; author via source)

Figure 2.6.5 Copy and Paste

Figure 2.6.6 Copy and Paste

- Terry's mean is 3.7, Davis' mean is 2.7, Maris' mean is 4.6.
- Terry's median is three, Davis' median is three. Maris' median is four.
- It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

? Exercise 2.6.1

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

Figure 2.6.7 This dot plot matches the supplied data. The plot uses a number line from 0 to 14. It shows two x's over 0, four x's over 1, three x's over 2, one x over 3, two x's over the number 4, 5, 6, and 9, and 1 x each over 10 and 14. There are no x's over the numbers 7, 8, 11, 12, and 13.

b.

The Ages Former U.S Presidents Died	
4	6 9
5	3 6 7 7 8
6	0 0 3 3 4 4 5 6 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3

Key: 8|0 means 80.

c.

Figure 2.6.8 This is a histogram titled Hours Spent Playing Video Games on Weekends. The x-axis shows the number of hours spent playing video games with bars showing values at intervals of 5. The y-axis shows the number of students. The first bar for 0 - 4.99 hours has a height of 2. The second bar from 5 - 9.99 has a height of 3. The third bar from 10 - 14.99 has a height of 4. The fourth bar from 15 - 19.99 has a height of 7. The fifth bar from 20 - 24.99 has a height of 9.

Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **left (or negative) skewed** distribution has a shape like Figure 2.6.2. A **right (or positive) skewed** distribution has a shape like Figure 2.6.3. A **symmetrical** distribution looks like Figure 2.6.1.

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

? Exercise 2.7.2

1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5

Answer

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

? Exercise 2.7.3

16; 17; 19; 22; 22; 22; 22; 22; 23

? Exercise 2.7.4

87; 87; 87; 87; 87; 88; 89; 89; 90; 91

Answer

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

? Exercise 2.7.5

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

? Exercise 2.7.6

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

Answer

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

? Exercise 2.7.7

What word describes a distribution that has two modes?

? Exercise 2.7.8

Describe the shape of this distribution.

Figure 2.6.9 This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right.

Answer

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

? Exercise 2.7.9

Describe the relationship between the mode and the median of this distribution.

Figure 2.6.10 This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heights from left to right are: 8, 4, 2, 2, 1.

? Exercise 2.7.10

Describe the relationship between the mean and the median of this distribution.

Figure 2.6.11 This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heights from left to right are: 8, 4, 2, 2, 1.

Answer

The mean is 4.1 and is slightly greater than the median, which is four.

? Exercise 2.7.11

Describe the shape of this distribution.

Figure 2.6.12

? Exercise 2.7.12

Describe the relationship between the mode and the median of this distribution.

Figure 2.6.13

Answer

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

? Exercise 2.7.13

Are the mean and the median the exact same in this distribution? Why or why not?

Figure 2.6.14

? Exercise 2.7.14

Describe the shape of this distribution.

Figure 2.6.15

Answer

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

? Exercise 2.7.15

Describe the relationship between the mode and the median of this distribution.

Figure 2.6.16 Copy and Paste Caption here. (Copyright; author via source)

? Exercise 2.7.16

Describe the relationship between the mean and the median of this distribution.

Figure 2.6.17

Answer

The mean and the median are both six.

? Exercise 2.7.17

The mean and median for the data are the same.

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

Is the data perfectly symmetrical? Why or why not?

? Exercise 2.7.18

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

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

Answer

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

? Exercise 2.7.19

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

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

? Exercise 2.7.20

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

Answer

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

? Exercise 2.7.21

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

This page titled [2.6: Skewness and the Mean, Median, and Mode](#) 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.

2.7: Measures of the Spread of the Data

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

The standard deviation

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

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

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

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

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

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

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

Rosa waits for seven minutes:

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

Binh waits for one minute.

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

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

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

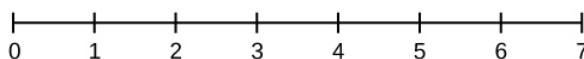


Figure 2.7.1

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer

- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

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

- sample:

$$x = \bar{x} + (\#ofSTDEV)(s) \quad (2.7.1)$$

- Population:

$$x = \mu + (\#ofSTDEV)(s) \quad (2.7.2)$$

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

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

Calculating the Standard Deviation

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

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

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

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

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad (2.7.3)$$

or

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \quad (2.7.4)$$

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

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \quad (2.7.5)$$

or

$$\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}} \quad (2.7.6)$$

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

In Equations 2.7.4 and 2.7.6, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Section 2.6. How much the statistic varies from one sample to another is known as the sampling variability of a statistic. You typically measure the **sampling variability of a statistic** by its standard error.

The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in Chapter 7. The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

✓ Example 2.7.1

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

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

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

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

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

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

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

$$s^2 = \frac{9.7375}{20-1} = 0.5125$$

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

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

and this is rounded to two decimal places, $s = 0.72$.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation on a calculator or computer.
- For a sample: $x = \bar{x} + (\text{\#ofSTDEVs})(s)$
- For a population: $x = \mu + (\text{\#ofSTDEVs})\sigma$
- For this example, use $x = \bar{x} + (\text{\#ofSTDEVs})(s)$ because the data is from a sample
 - a. Verify the mean and standard deviation on your calculator or computer.
 - b. Find the value that is one standard deviation above the mean. Find $(\bar{x} + 1s)$.
 - c. Find the value that is two standard deviations below the mean. Find $(\bar{x} - 2s)$.
 - d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

- a.
 - o Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
 - o Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
 - o Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
 - o Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
 - o $\bar{x} = 10.525$
 - o Use S_x because this is sample data (not a population): $S_x = 0.715891$
- b. $(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$
- c. $(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$
- d.
 - o $(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
 - o $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

? Exercise 2.7.2

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

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

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Answer

$$\bar{x} = 30.68$$

$$s = 6.09$$

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

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the

mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is -1.525 for the data value nine. **If you add the deviations, the sum is always zero.** (For Example 2.7.1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

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

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

Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

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

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

✓ Example 2.7.3

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

- a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
 - i. The sample mean
 - ii. The sample standard deviation
 - iii. The median
 - iv. The first quartile
 - v. The third quartile
 - vi. *IQR*
- c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Answer

- a. See Table
- b.
 - i. The sample mean = 73.5
 - ii. The sample standard deviation = 17.9
 - iii. The median = 73
 - iv. The first quartile = 61
 - v. The third quartile = 90
 - vi. *IQR* = $90 - 61 = 29$
- c. The x -axis goes from 32.5 to 100.5; y -axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$

$86.9 + 13.6 = 100.5$ =the ending value; No data values fall on an interval boundary.

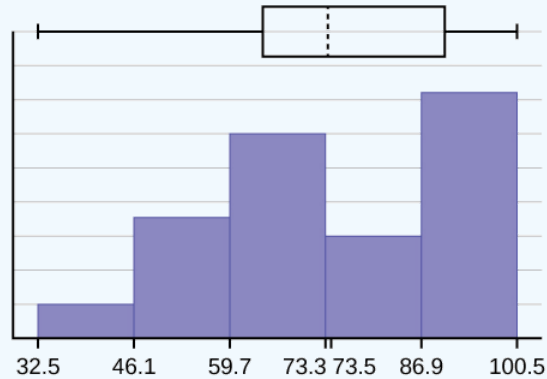


Figure 2.7.2.

The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ($73 - 33 = 40$) than the spread in the upper 50% ($100 - 73 = 27$). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores ($IQR = 29$) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

? Exercise 2.7.4

The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Answer

$$\bar{x} = 9.3 \text{ and } s = 2.2$$

Comparing Values from Different Data Sets

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

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

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

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

✓ Example 2.7.5

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

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Answer

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

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

For John,

$$z = \text{\#ofSTDEVs} = \left(\frac{2.85 - 3.0}{0.7} \right) = -0.21$$

For Ali,

$$z = \frac{\text{GPA} - \text{Mean GPA}}{\text{STDEVs}} = \left(\frac{77 - 80}{10} \right) = -0.3$$

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

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

? Exercise 2.7.6

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

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Answer

For Angie:

$$z = \left(\frac{26.2 - 27.2}{0.8} \right) = -1.25$$

For Beth:

$$z = \left(\frac{27.3 - 30.1}{1.4} \right) = -2$$

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

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

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

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

References

1. Data from Microsoft Bookshelf.
2. King, Bill. "Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ , and the formula $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$.

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \quad (2.7.7)$$

where s_x = sample standard deviation and $\bar{x}$ = sample mean

Glossary

Standard Deviation

Contributors and Attributions

Variance

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

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

? Exercise 2.7.7

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Answer

$$s = 34.5$$

? Exercise 2.7.8

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

? Exercise 2.7.9

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Answer

For Fredo:

$$z = \frac{0.158 - 0.166}{0.012} = -0.67$$

For Karl:

$$z = \frac{0.177 - 0.189}{0.015} = -0.8$$

Fredo's z-score of -0.67 is higher than Karl's z-score of -0.8 . For batting average, higher values are better, so Fredo has a better batting average compared to his team.

? Exercise 2.7.10

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; -1 ; 2; 0; 5; -3 ; 1; -1 ; 6; 5; -2

- What is the mean change score?
- What is the standard deviation for this population?
- What is the median change score?
- Find the change score that is 2.2 standard deviations below the mean.

? Exercise 2.7.11

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

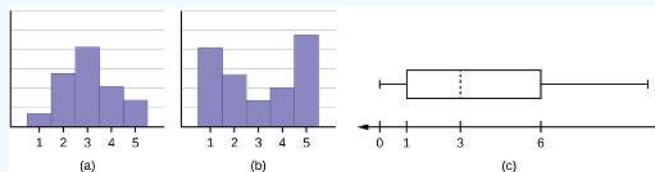


Figure 2.7.6.

- The medians for all three graphs are the same.
- We cannot determine if any of the means for the three graphs is different.
- The standard deviation for graph b is larger than the standard deviation for graph a.
- We cannot determine if any of the third quartiles for the three graphs is different.

Answer

- True
- True
- True
- False

This page titled [2.7: Measures of the Spread of the Data](#) 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.

- 2.8: Measures of the Spread of the Data** by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

CHAPTER OVERVIEW

3: 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.

[3.0: Prelude to Linear Regression and Correlation](#)

[3.1: Linear Equations](#)

[3.2: Scatter Plots](#)

[3.3: The Regression Equation](#)

[3.4: Prediction](#)

This page titled [3: 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.

3.0: Prelude 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 3.0.1: 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 [3.0: Prelude 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.

3.1: 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 3.1.1

The following examples are linear equations.

$$y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

? Exercise 3.1.1

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 3.1.2

Graph the equation $y = -1 + 2x$.

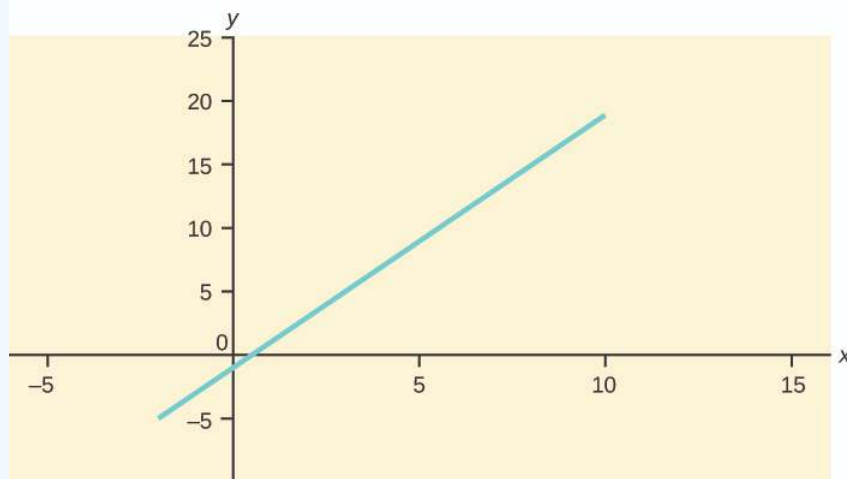


Figure 3.1.1

? Exercise 3.1.2

Is the following an example of a linear equation? Why or why not?

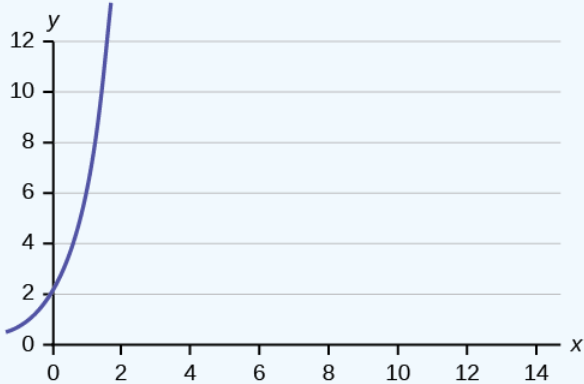


Figure 3.1.2

Answer

No, the graph is not a straight line; therefore, it is not a linear equation.

✓ **Example 3.1.3**

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 3.1.3**

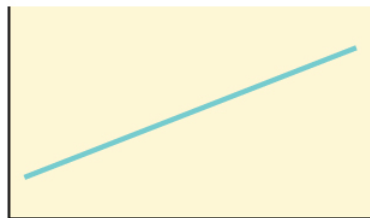
Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of \$50 per class as well as \$20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Answer

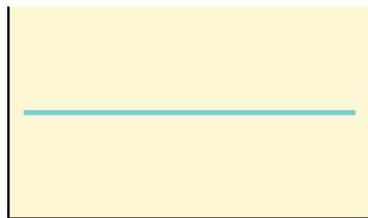
$$y = 50 + 20x$$

Slope and Y-Intercept of a Linear Equation

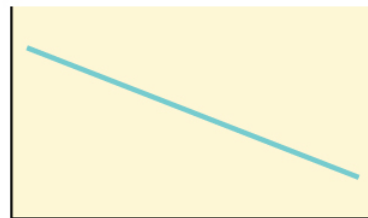
For the linear equation $y = a + bx$, b = slope and a = y -intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y -intercept is the y coordinate of the point $(0, a)$ where the line crosses the y -axis.



(a)



(b)



(c)

Figure 3.1.3.

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 3.1.4

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 3.1.4

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 [3.1: 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.

3.2: 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 3.2.1

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 3.2.1: 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

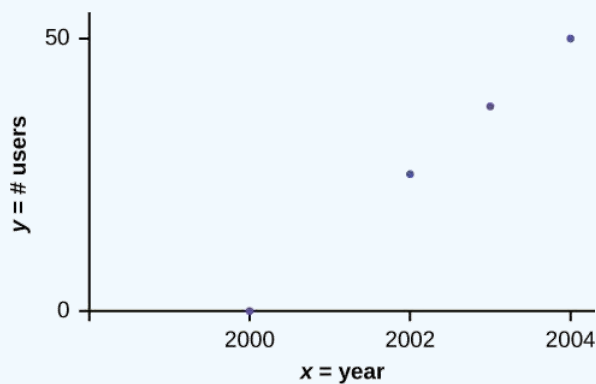


Figure 3.2.1: Scatter plot showing the number of m-commerce users

(in millions) by year.

📌 To create a scatter plot

- Enter your X data into list L1 and your Y data into list L2.
- 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.)
- For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
- For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
- For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
- Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
- 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 3.2.1

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

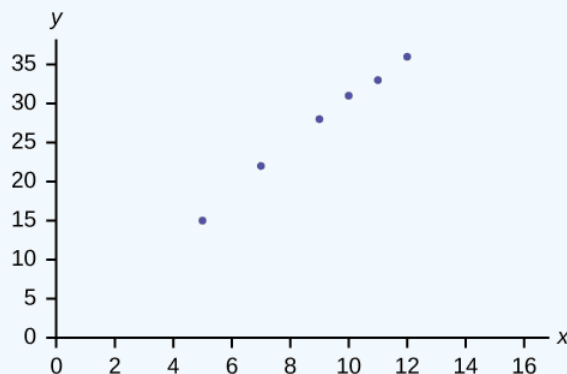


Figure 3.2.2

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.

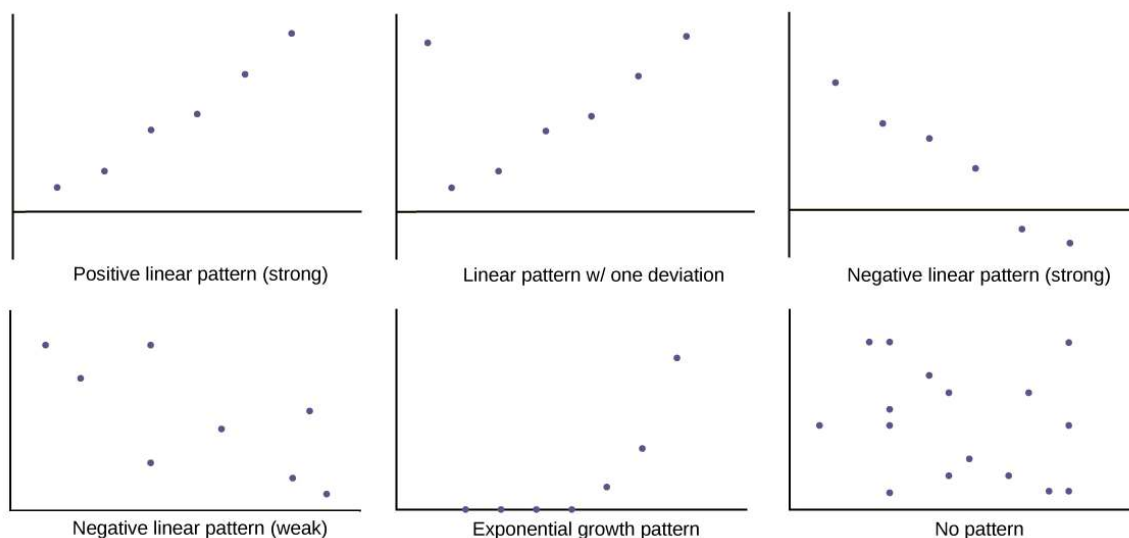


Figure 3.2.3

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 [3.2: 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.

3.3: The Regression Equation

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 3.3.1

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

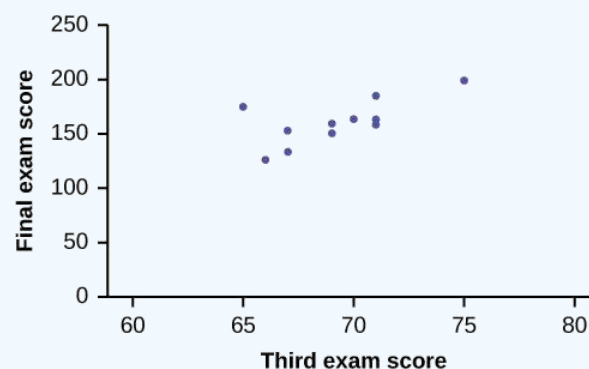


Figure 3.3.1: Scatter plot showing the scores on the final exam based on scores from the third exam.

? Exercise 3.3.1

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 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.

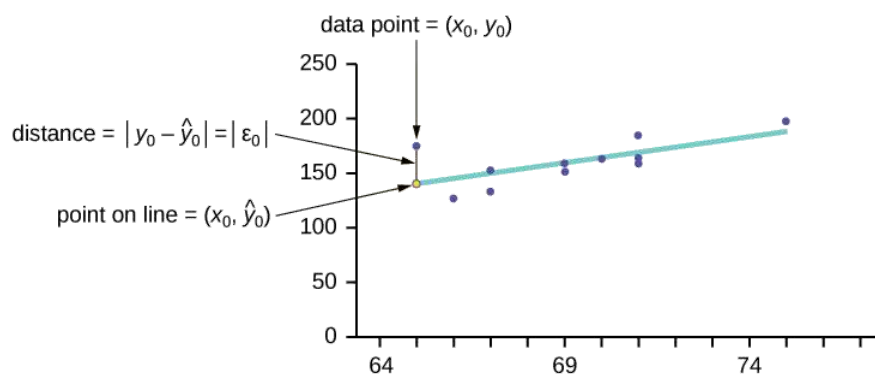


Figure 3.3.2

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.

ε = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, \dots, 11$.

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

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + \dots + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2 \quad (3.3.1)$$

Equation 3.3.1 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:

$$\hat{y} = a + bx \quad (3.3.2)$$

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 y values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r \left(\frac{s_y}{s_x} \right)$ where s_y = 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 below.

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:

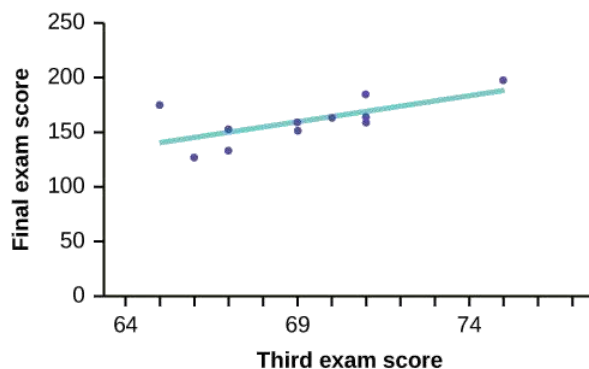


Figure 3.3.3

The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

$$\hat{y} = -173.51 + 4.83x \quad (3.3.3)$$

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

- 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.)
- 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.)
- On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
- On the next line, at the prompt β or ρ , highlight " $\neq 0$ " and press ENTER
- Leave the line for "RegEq:" blank
- Highlight Calculate and press ENTER.

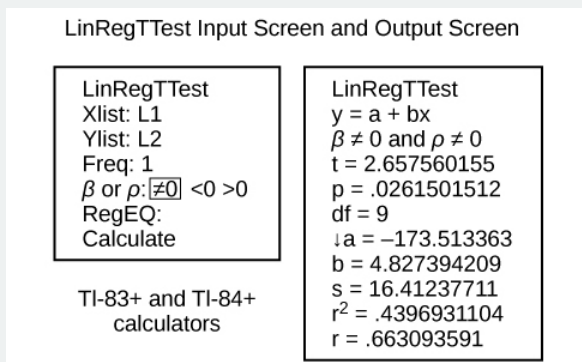


Figure 3.3.4

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

- We are assuming your X data is already entered in list L1 and your Y data is in list L2
- 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

$$r = \frac{n \sum(xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (3.3.4)$$

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**."

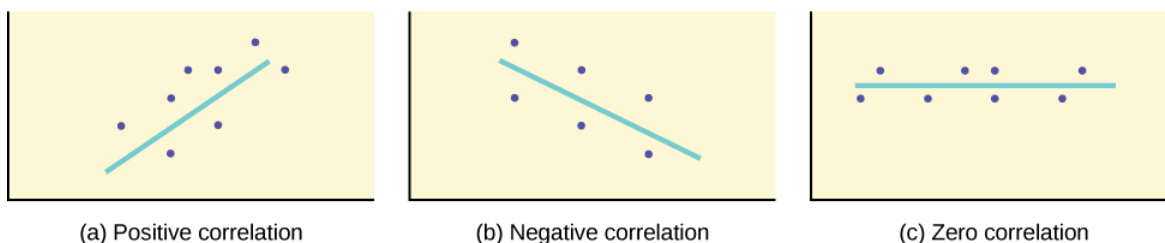


Figure 3.3.5: (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

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:

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{\left[n \sum x^2 - (\sum x)^2\right] \left[n \sum y^2 - (\sum y)^2\right]}} \quad (3.3.5)$$

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 [3.3: The Regression Equation](#) 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: 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 3.4.1

Recall the third exam/final exam example.

- What would you predict the final exam score to be for a student who scored a 66 on the third exam?
- 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 3.4.1

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

- Data from the Centers for Disease Control and Prevention.
- 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 [3.4: 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.

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.

[4.0: Introduction](#)

[4.1: Terminology](#)

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.

4.0: Introduction

 This is a photo taken of the night sky. A meteor and its tail are shown entering the earth's atmosphere. **Figure 4.0.1** 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 [4.0: 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: 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 $A \text{ OR } B$ if the outcome is in A or is in B or is in both A and B . For example, let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. $A \text{ OR } B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

The "AND" Event

An outcome is in the event $A \text{ AND } B$ 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 $A \text{ AND } B = \{4, 5\}$.

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

$$P(A) + P(A') = 1.$$

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

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

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

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. $P(A) =$ _____, $P(B) =$ _____

d. $A \text{ AND } B =$ _____, $A \text{ OR } B =$ _____

e. $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$ _____

f. $A' =$ _____, $P(A') =$ _____

g. $P(A) + P(A') =$ _____

Answer

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$

b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$

c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$

d. $A \text{ AND } B = \{14, 16, 18\}$, $A \text{ OR } B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$

e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$

f. $A' = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$, $P(A') = \frac{10}{19}$

g. $P(A) + P(A') = 1 \left(\left(\frac{9}{19} + \frac{10}{19} \right) = 1 \right)$

? Exercise 4.1.1

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)).

- $S =$ _____
- Let event A = the sum is even and event B = the first number is prime.
- $A =$ _____, $B =$ _____
- $P(A) =$ _____, $P(B) =$ _____
- $A \text{ AND } B =$ _____, $A \text{ OR } B =$ _____
- $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$ _____
- $B' =$ _____, $P(B') =$ _____
- $P(A) + P(A') =$ _____

Answer

- $S = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4)\}$
- $A = \{(1, 1), (1, 3), (2, 2), (2, 4), (3, 1), (3, 3)\}$
 $B = \{(2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4)\}$
- $P(A) = \frac{1}{2}$, $P(B) = \frac{2}{3}$
- $A \text{ AND } B = \{(2, 2), (2, 4), (3, 1), (3, 3)\}$
 $A \text{ OR } B = \{(1, 1), (1, 3), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4)\}$
- $P(A \text{ AND } B) = \frac{1}{3}$, $P(A \text{ OR } B) = \frac{5}{6}$
- $B' = \{(1, 1), (1, 2), (1, 3), (1, 4)\}$, $P(B') = \frac{1}{3}$
- $P(B) + P(B') = 1$

✓ Example 4.1.2

A fair, six-sided die is rolled. Describe the sample space S , identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

- Event T = the outcome is two.
- Event A = the outcome is an even number.
- Event B = the outcome is less than four.
- The complement of A .
- $A \text{ GIVEN } B$
- $B \text{ GIVEN } A$
- $A \text{ AND } B$
- $A \text{ OR } B$
- $A \text{ OR } B'$
- Event N = the outcome is a prime number.
- Event I = the outcome is seven.

Solution

- $T = \{2\}$, $P(T) = \frac{1}{6}$
- $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$
- $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$
- $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$
- $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$
- $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$
- $A \text{ AND } B = 2$, $P(A \text{ AND } B) = \frac{1}{6}$
- $A \text{ OR } B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = \frac{5}{6}$
- $A \text{ OR } B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = \frac{2}{3}$
- $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$

k. A six-sided die does not have seven dots. $P(7) = 0$.

? Exercise 4.1.2

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:

- $P(M)$
- $P(F)$
- $P(R)$
- $P(L)$
- $P(M \text{ AND } R)$
- $P(F \text{ AND } L)$
- $P(M \text{ OR } F)$
- $P(M \text{ OR } R)$
- $P(F \text{ OR } L)$
- $P(M')$

Answer

- $P(M) = 0.52$
- $P(F) = 0.48$
- $P(R) = 0.87$
- $P(L) = 0.13$
- $P(M \text{ AND } R) = 0.43$
- $P(F \text{ AND } L) = 0.04$
- $P(M \text{ OR } F) = 1$
- $P(M \text{ OR } R) = 0.96$
- $P(F \text{ OR } L) = 0.57$
- $P(M') = 0.48$

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

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

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

Glossary

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:

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

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event $A \text{ AND } B$ if the outcome is in both $A \text{ AND } B$ at the same time.

The Complement Event

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

The Or Event

An outcome is in the event $A \text{ OR } B$ if the outcome is in A or is in B or is in both A and B .

This page titled [4.1: 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.

CHAPTER OVERVIEW

5: Discrete Random Variables

[5.0: Prelude to Discrete Random Variables](#)

[5.1: Probability Distribution Function \(PDF\) for a Discrete Random Variable](#)

[5.2: Mean and Standard Deviation](#)

[5.3: Binomial Distribution](#)

This page titled [5: Discrete Random Variables](#) 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.

5.0: Prelude to Discrete Random Variables

CHAPTER OBJECTIVES

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

- Recognize and understand discrete probability distribution functions, in general.
 - Calculate and interpret expected values.
 - Recognize the binomial probability distribution and apply it appropriately.
 - Recognize the Poisson probability distribution and apply it appropriately.
 - Recognize the geometric probability distribution and apply it appropriately.
 - Recognize the hypergeometric probability distribution and apply it appropriately.
 - Classify discrete word problems by their distributions.
- A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?
 - Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A *random variable* describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.



Figure 5.0.1 You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Random Variable Notation

Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. **If X is a random variable, then X is written in words, and x is given as a number.**

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is TTT ; THH ; HTH ; HHT ; HTT ; THT ; TTH ; HHH . Then, $x = 0, 1, 2, 3$. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Collaborative Exercise

Toss a coin ten times and record the number of heads. After all members of the class have completed the experiment (tossed a coin ten times and counted the number of heads), fill in [Table](#). Let X = the number of heads in ten tosses of the coin.

x	Frequency of x	Relative Frequency of x

x	Frequency of x	Relative Frequency of x

- Which value(s) of x occurred most frequently?
- If you tossed the coin 1,000 times, what values could x take on? Which value(s) of x do you think would occur most frequently?
- What does the relative frequency column sum to?

Glossary

Random Variable (RV)

a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters $X, Y, Z, \dots$; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y , and z . For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if $X = \text{hair color}$ then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x the random variable X takes only after performing the experiment.

This page titled [5.0: Prelude to Discrete Random Variables](#) 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.

5.1: Probability Distribution Function (PDF) for a Discrete Random Variable

A discrete probability distribution function has two characteristics:

- Each probability is between zero and one, inclusive. (*inclusive* means to include zero and one).
- The sum of the probabilities is one.

✓ Example 5.1.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$

$P(x)$ = probability that X takes on a value x .

x	$P(x)$
0	$P(x = 0) = \frac{2}{50}$
1	$P(x = 1) = \frac{11}{50}$
2	$P(x = 2) = \frac{23}{50}$
3	$P(x = 3) = \frac{9}{50}$
4	$P(x = 4) = \frac{4}{50}$
5	$P(x = 5) = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because:

- Each $P(x)$ is between zero and one, inclusive.
- The sum of the probabilities is one, that is,

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1 \quad (5.1.1)$$

? Exercise 5.1.1

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, $x = 0, 1, 2, 3, 4, 5$ $P(x)$ = the probability that X takes on value x . Why is this a discrete probability distribution function (two reasons)?

X	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$

X	$P(x)$
5	$P(x = 5) = \frac{2}{50}$

Answer

Each $P(x)$ is between 0 and 1, inclusive, and the sum of the probabilities is 1, that is:

$$\frac{4}{50} + \frac{8}{50} + \frac{16}{50} + \frac{14}{50} + \frac{6}{50} + \frac{2}{50} = 1 \quad (5.1.2)$$

✓ Example 5.1.2

Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

- Let X = the number of days Nancy _____.
- X takes on what values?
- Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in [Example](#). The table should have two columns labeled x and $P(x)$. What does the $P(x)$ column sum to?

Solutions

- Let X = the number of days Nancy attends class per week.
- 0, 1, 2, and 3
-

x	$P(x)$
0	0.01
1	0.04
2	0.15
3	0.80

? Exercise 5.1.2

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

Answer

X is the number of days Jeremiah attends basketball practice per week. X takes on the values 0, 1, and 2.

Use the following information to answer the next three exercises: A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution.

Let X = the number of years a new hire will stay with the company.

Let $P(x)$ = the probability that a new hire will stay with the company x years.

? Exercise 5.1.3

Complete the table below using the data provided.

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

Answer

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

? Exercise 5.1.4

$$P(x \geq 5) = \underline{\hspace{2cm}}$$

Answer

$$0.10 + 0.05 = 0.15$$

? Exercise 5.1.5

What does the column " $P(x)$ " sum to?

Answer

1

Use the following information to answer the next six exercises: A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	0.15

x	$P(x)$
2	0.35
3	0.40
4	0.10

? Exercise 5.1.6

What is the probability the baker will sell more than one batch? $P(x > 1) = \underline{\hspace{2cm}}$

Answer

$$0.35 + 0.40 + 0.10 = 0.85$$

? Exercise 5.1.7

On average, how many batches should the baker make?

Answer

$$1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45$$

Use the following information to answer the next exercise: Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

? Exercise 5.1.8

Construct a probability distribution table for the data.

Answer

x	$P(x)$
0	0.03
1	0.04
2	0.08
3	0.85

Use the following information to answer the next three exercises: Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

? Exercise 5.1.9

Define the random variable X .

Answer

Let X = the number of events Javier volunteers for each month.

? Exercise 5.1.10

Construct a PDF table.

Answer

x	$P(x)$
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

? Exercise 5.1.11

Find the probability that Javier volunteers for at least one event each month. $P(x > 0) = \underline{\hspace{2cm}}$

Answer

$$1 - 0.05 = 0.95$$

Glossary

Probability Distribution Function (PDF)

a mathematical description of a discrete random variable (RV), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

This page titled [5.1: Probability Distribution Function \(PDF\) for a Discrete Random Variable](#) 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.

5.2: Mean and Standard Deviation

The expected value is often referred to as the "long-term" average or mean. This means that over the long term of doing an experiment over and over, you would expect this average.

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. As you learned in Chapter 3, probability does not describe the short-term results of an experiment. It gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the Law of Large Numbers.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero (the theoretical probability and the relative frequency get closer and closer together). When evaluating the long-term results of statistical experiments, we often want to know the "average" outcome. This "long-term average" is known as the mean or expected value of the experiment and is denoted by the Greek letter μ . In other words, after conducting many trials of an experiment, you would expect this average value.

To find the expected value or long term average, μ , simply multiply each value of the random variable by its probability and add the products.

✓ Example 5.2.1

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ , of the number of days per week the men's soccer team plays soccer.

Solution

To do the problem, first let the random variable X = the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table adding a column $x * P(x)$. In this column, you will multiply each x value by its probability.

Expected Value Table This table is called an expected value table. The table helps you calculate the expected value or long-term average.

x	$P(x)$	$x * P(x)$
0	0.2	$(0)(0.2) = 0$
1	0.5	$(1)(0.5) = 0.5$
2	0.3	$(2)(0.3) = 0.6$

Add the last column $x * P(x)$ to find the long term average or expected value:

$$(0)(0.2) + (1)(0.5) + (2)(0.3) = 0 + 0.5 + 0.6 = 1.1.$$

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week. We say $\mu = 1.1$.

✓ Example 5.2.2

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

You expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

x	$P(x)$	$x * P(x)$	$(x - \mu)^2 * P(x)$
0	$P(x = 0) = \frac{2}{50}$	$(0) \left(\frac{2}{50} \right) = 0$	$(0 - 2.1)^2 \cdot 0.04 = 0.1764$

x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 \cdot P(x)$
1	$P(x = 1) = \frac{11}{50}$	$(1) \left(\frac{11}{50} \right) = \frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$
2	$P(x = 2) = \frac{23}{50}$	$(2) \left(\frac{23}{50} \right) = \frac{46}{50}$	$(2 - 2.1)^2 \cdot 0.46 = 0.0046$
3	$P(x = 3) = \frac{9}{50}$	$(3) \left(\frac{9}{50} \right) = \frac{27}{50}$	$(3 - 2.1)^2 \cdot 0.18 = 0.1458$
4	$P(x = 4) = \frac{4}{50}$	$(4) \left(\frac{4}{50} \right) = \frac{16}{50}$	$(4 - 2.1)^2 \cdot 0.08 = 0.2888$
5	$P(x = 5) = \frac{1}{50}$	$(5) \left(\frac{1}{50} \right) = \frac{5}{50}$	$(5 - 2.1)^2 \cdot 0.02 = 0.1682$

Add the values in the third column of the table to find the expected value of X :

$$\mu = \text{Expected Value} = \frac{105}{50} = 2.1$$

Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x , multiply the square of its deviation by its probability. (Each deviation has the format $x - \mu$.)

Add the values in the fourth column of the table:

$$0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05$$

The standard deviation of X is the square root of this sum: $\sigma = \sqrt{1.05} \approx 1.0247$

The mean, μ , of a discrete probability function is the expected value.

$$\mu = \sum (x \cdot P(x))$$

The standard deviation, Σ , of the PDF is the square root of the variance.

$$\sigma = \sqrt{\sum [(x - \mu)^2 \cdot P(x)]}$$

When all outcomes in the probability distribution are equally likely, these formulas coincide with the mean and standard deviation of the set of possible outcomes.

? Exercise 5.2.2

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

x	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Answer

The expected value is 2.24

$$(0)\frac{4}{50} + (1)\frac{8}{50} + (2)\frac{16}{50} + (3)\frac{14}{50} + (4)\frac{6}{50} + (5)\frac{2}{50} = 0 + \frac{8}{50} + \frac{32}{50} + \frac{42}{50} + \frac{24}{50} + \frac{10}{50} = \frac{116}{50} = 2.32 \quad (5.2.1)$$

✓ Example 5.2.2

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay \$2 to play and could profit \$100,000 if you match all five numbers in order (you get your \$2 back plus \$100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let X = the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and -2 dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is

$$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = (1)(10^{-5}) \\ = 0.00001.$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

$$1 - 0.00001 = 0.99999. 1 - 0.00001 = 0.99999.$$

The expected value table is as follows:

Add the last column. $-1.99998 + 1 = -0.99998$

	x	$P(x)$	$xP(x)$
Loss	-2	0.99999	$(-2)(0.99999) = -1.99998$
Profit	100,000	0.00001	$(100000)(0.00001) = 1$

Since -0.99998 is about -1, you would, on average, expect to lose approximately \$1 for each game you play. However, each time you play, you either lose \$2 or profit \$100,000. The \$1 is the average or expected LOSS per game after playing this game over and over.

? Exercise 5.2.3

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay \$1 to play. If you guess the right suit every time, you get your money back and \$256. What is your expected profit of playing the game over the long term?

Answer

Let X = the amount of money you profit. The x -values are -\$1 and \$256.

The probability of guessing the right suit each time is $\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{256} = 0.0039$

The probability of losing is $1 - \frac{1}{256} = \frac{255}{256} = 0.9961$

$$(0.0039)256 + (0.9961)(-1) = 0.9984 + (-0.9961) = 0.0023 \text{ or } 0.23\text{cents}.$$

✓ Example 5.2.4

Suppose you play a game with a biased coin. You play each game by tossing the coin once. $P(\text{heads}) = \frac{2}{3}$ and $P(\text{tails}) = \frac{1}{3}$. If you toss a head, you pay \$6. If you toss a tail, you win \$10. If you play this game many times, will you come out ahead?

- Define a random variable X .
- Complete the following expected value table.
- What is the expected value, μ ? Do you come out ahead?

Solutions

a.

X = amount of profit

	x		
WIN	10	$\frac{1}{3}$	
LOSE			$-\frac{12}{3}$

b.

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	-6	$\frac{2}{3}$	$-\frac{12}{3}$

c.

Add the last column of the table. The expected value $\mu = -\frac{2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

? Exercise 5.2.4

Suppose you play a game with a spinner. You play each game by spinning the spinner once. $P(\text{red}) = \frac{2}{5}$, $P(\text{blue}) = \frac{2}{5}$, and $P(\text{green}) = \frac{1}{5}$. If you land on red, you pay \$10. If you land on blue, you don't pay or win anything. If you land on green, you win \$10. Complete the following expected value table.

	x	$P(x)$	
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Answer

	x	$P(x)$	$x * P(x)$
Red	-10	$\frac{2}{5}$	$-\frac{20}{5}$
Blue	0	$\frac{2}{5}$	0
Green	10	$\frac{1}{5}$	$\frac{10}{5}$

	x	$P(x)$	$x * P(x)$
Blue	0	$\frac{2}{5}$	$\frac{0}{5}$
Green	10	$\frac{1}{5}$	$\frac{1}{5}$

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root. To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x - \mu)^2 P(x)$ and take the square root.

x	$P(x)$	$x * P(x)$	$(x - \mu)^2 P(x)$
0	0.2	$(0)(0.2) = 0$	$(0 - 1.1)^2(0.2) = 0.242$
1	0.5	$(1)(0.5) = 0.5$	$(1 - 1.1)^2(0.5) = 0.005$
2	0.3	$(2)(0.3) = 0.6$	$(2 - 1.1)^2(0.3) = 0.243$

Add the last column in the table. $0.242 + 0.005 + 0.243 = 0.490$ The standard deviation is the square root of 0.49, or $\sigma = \sqrt{0.49} = 0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas for calculating μ and σ .

✓ Example 5.2.5

Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like Table and calculate the mean μ and standard deviation σ of X .

Solution

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Use the sample space to complete the following table:

Calculating μ and σ .

x	$P(x)$	$xP(x)$	$(x - \mu)^2 \cdot P(x)$
0	$\frac{9}{36}$	0	$(0 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1 - 1)^2 \cdot \frac{18}{36} = 0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(2 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$

Add the values in the third column to find the expected value: $\mu = \frac{36}{36} = 1$. Use this value to complete the fourth column.

Add the values in the fourth column and take the square root of the sum:

$$\sigma = \sqrt{\frac{18}{36}} \approx 0.7071. \quad (5.2.2)$$

✓ Exercise 5.2.5

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win \$50. If you lose the bet, you pay \$20. Let X = the amount of profit from a bet.

$$P(\text{win}) = P(\text{one moderate earthquake will occur}) = 0.2142$$

$$P(\text{loss}) = P(\text{one moderate earthquake will not occur}) = 0.7858$$

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of X ? Construct a table similar to [Table 4.12](#) and [Table 4.13](#) to help you answer these questions.

Answer

	x	$P(x)$	$xP(x)$	$(x - \mu^2)P(x)$
win	50	0.2142	10.71	$[50 - (-5.006)]^2(0.2142)$ $= 648.0964$
loss	-20	0.7858	-15.716	$[-20 - (-5.006)]^2(0.7858) =$ 176.6636

$$\text{Mean} = \text{Expected Value} = 10.71 + (-15.716) = -5.006.$$

If you make this bet many times under the same conditions, your long term outcome will be an average *loss* of \$5.01 per bet.

$$\text{Standard Deviation} = \sqrt{648.0964 + 176.6636} \approx 28.7186$$

? Exercise 5.2.6

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. You bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win \$100. If you lose the bet, you pay \$10. Let X = the amount of profit from a bet. Find the mean and standard deviation of X .

Answer

	x	$P(x)$	$x \cdot P(x)$	$(x - \mu^2) \cdot P(x)$
win	100	0.0108	1.08	$[100 - (-8.812)]^2 \cdot$ $0.0108 = 127.8726$
loss	-10	0.9892	-9.892	$[-10 - (-8.812)]^2 \cdot$ $0.9892 = 1.3961$

$$\text{Mean} = \text{Expected Value} = \mu = 1.08 + (-9.892) = -8.812$$

If you make this bet many times under the same conditions, your long term outcome will be an average *loss* of \$8.81 per bet.

$$\text{Standard Deviation} = \sqrt{127.8726 + 1.3961} \approx 11.3696$$

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

Summary

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

Formula Review

1. Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$
2. Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$

Glossary

Expected Value

expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: μ . For a discrete random variable (RV) with probability distribution function $P(x)$, the definition can also be written in the form $\mu = \sum xP(x)$.

Mean

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

Mean of a Probability Distribution

the long-term average of many trials of a statistical experiment

Standard Deviation of a Probability Distribution

a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution

The Law of Large Numbers

As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.

References

1. Class Catalogue at the Florida State University. Available online at apps.oti.fsu.edu/RegistrarCo...archFormLegacy (accessed May 15, 2013).
2. "World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. www.world-earthquakes.com/ind...thq_prediction (accessed May 15, 2013).

This page titled [5.2: Mean and Standard Deviation](#) 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.

- [4.3: Mean or Expected Value and Standard Deviation](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

5.3: Binomial Distribution

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn *with replacement* from a population of size N .

Three characteristics of a binomial experiment

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p , of a success and probability, q , of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X = the number of successes obtained in the n independent trials. The mean, μ , and variance, σ^2 , for the binomial probability distribution are

$$\mu = np \quad (5.3.1)$$

and

$$\sigma^2 = npq. \quad (5.3.2)$$

The standard deviation, σ , is then

$$\sigma = \sqrt{npq}. \quad (5.3.3)$$

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example 5.3.1

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable X = the number of students who withdraw from the randomly selected elementary physics class.

Exercise 5.3.1

The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a "success" be in this case?

Answer

A school that offers fruit in their lunch every day

Example 5.3.2

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times.

Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

? Exercise 5.3.2

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

Answer

$$P(x = 12)$$

✓ Example 5.3.3

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically.

Solution

$$P(x > 10)$$

? Exercise 5.3.3

A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

Answer

$$P(x > 3)$$

✓ Example 5.3.4

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

- This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.
- If we are interested in the number of students who do their homework on time, then how do we define X ?
- What values does x take on?
- What is a "failure," in words?
- If $p + q = 1$, then what is q ?
- The words "at least" translate as what kind of inequality for the probability question $P(x \text{ ____ } 40)$.

Solution

- failure
- X = the number of statistics students who do their homework on time
- 0, 1, 2, ..., 50
- Failure is defined as a student who does not complete his or her homework on time. The probability of a success is $p = 0.70$. The number of trials is $n = 50$.
- $q = 0.30$
- greater than or equal to ($\geq$). The probability question is $P(x \geq 40)$.

? Exercise 5.3.4

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Give two reasons why this is a binomial problem.

Answer

This is a binomial problem because there is only a success or a failure, and there are a definite number of trials. The probability of a success stays the same for each trial.

📌 Notation for the Binomial: $B =$ Binomial Probability Distribution Function

$$X \sim B(n, p) \quad (5.3.4)$$

Read this as " X is a random variable with a binomial distribution." The parameters are n and p ; n = number of trials, p = probability of a success on each trial.

✓ Example 5.3.5

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20$, $p = 0.41$, and $q = 1 - 0.41 = 0.59$. $X \sim B(20, 0.41)$

Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$. (calculator or computer)

Go into 2nd DISTR. The syntax for the instructions are as follows:

To calculate (x = value) : binompdf(n, p , number) if "number" is left out, the result is the binomial probability table.

To calculate $P(x \leq \text{value})$: binomcdf(n, p , number) if "number" is left out, the result is the cumulative binomial probability table.

For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,0.41,12). The result is $P(x \leq 12) = 0.9738$.

If you want to find $P(x = 12)$, use binompdf(20,0.41,12) If you want to find $P(x > 12)$, use $1 - \text{binomcdf}(20, 0.41, 12)$.

The probability that at most 12 workers have a high school diploma but do not pursue any further education is 0.9738.

The graph of $X \sim B(20, 0.41)$ is as follows:

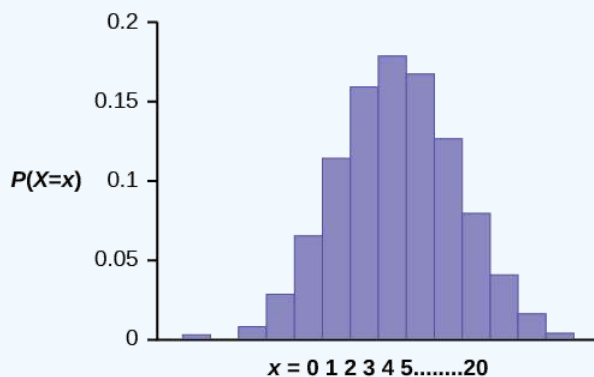


Figure 5.3.1 : The graph of $X \sim B(20, 0.41)$.

The y -axis contains the probability of x , where X = the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)(0.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$.

$$\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20. \quad (5.3.5)$$

? Exercise 5.3.5

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

- What values does x take on?
- What is the probability distribution? Find the following probabilities:
 - the probability that two pages feature signature artists
 - the probability that at most six pages feature signature artists
 - the probability that more than three pages feature signature artists.
- Using the formulas, calculate the (i) mean and (ii) standard deviation.

Answer

- $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
- $X \sim B(100, 8/560)$
 - $P(x = 2) = \text{binompdf}\left(100, \frac{8}{560}, 2\right) = 0.2466$
 - $P(x \leq 6) = \text{binomcdf}\left(100, \frac{8}{560}, 6\right) = 0.9994$
 - $P(x > 3) = 1 - P(x \leq 3) = 1 - \text{binomcdf}\left(100, \frac{8}{560}, 3\right) = 1 - 0.9443 = 0.0557$
- Mean $= np = (100)\left(\frac{8}{560}\right) = \frac{800}{560} \approx 1.4286$
 - Standard Deviation $= \sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.1867$

? Exercise 5.3.6

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

- What is the probability distribution for X ?
- Use your calculator to find the following probabilities:
 - the probability that 25 adults in the sample prefer saving over spending
 - the probability that at most 20 adults prefer saving
 - the probability that more than 30 adults prefer saving
- Using the formulas, calculate the (i) mean and (ii) standard deviation of X .

Answer

- $X \sim B(50, 0.6)$
- Using the TI-83, 83+, 84 calculator with instructions as provided in [Example 4.13](#):
 - $P(x = 25) = \text{binompdf}(50, 0.6, 25) = 0.0405$
 - $P(x \leq 20) = \text{binomcdf}(50, 0.6, 20) = 0.0034$
 - $P(x > 30) = 1 - \text{binomcdf}(50, 0.6, 30) = 1 - 0.5535 = 0.4465$

- c. i. Mean = $np = 50(0.6) = 30$
- ii. Standard Deviation = $\sqrt{npq} = \sqrt{50(0.6)(0.4)} \approx 3.4641$

? Exercise 5.3.7

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

- a. What is the probability distribution for X ?
- b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X .
- c. Use your calculator to find the probability that at most eight people develop pancreatic cancer
- d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

Answer

- a. $X \sim B(200, 0.0128)$
- b. i. Mean = $np = 200(0.0128) = 2.56$
- ii. Standard Deviation = $\sqrt{npq} = \sqrt{(200)(0.0128)(0.9872)} \approx 1.5897$
- c. Using the TI-83, 83+, 84 calculator with instructions as provided in [Example 4.13](#):
 $P(x \leq 8) = \text{binomcdf}(200, 0.0128, 8) = 0.9988$
- d. $P(x = 5) = \text{binompdf}(200, 0.0128, 5) = 0.0707$
 $P(x = 6) = \text{binompdf}(200, 0.0128, 6) = 0.0298$
 So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop cancer than six.

? Exercise 5.3.8

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

- a. What is the probability distribution for X ?
- b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X .
- c. Use your calculator to find the probability that DeAndre scored with 60 of these shots.
- d. Find the probability that DeAndre scored with more than 50 of these shots.

Answer

- a. $X \sim B(80, 0.613)$
- b. i. Mean = $np = 80(0.613) = 49.04$
- ii. Standard Deviation = $\sqrt{npq} = \sqrt{80(0.613)(0.387)} \approx 4.3564$
- c. Using the TI-83, 83+, 84 calculator with instructions as provided in [Example 4.13](#):
 $P(x = 60) = \text{binompdf}(80, 0.613, 60) = 0.0036$
- d. $P(x > 50) = 1 - P(x \leq 50) = 1 - \text{binomcdf}(80, 0.613, 50) = 1 - 0.6282 = 0.3718$

✓ Example 5.3.9

The following example illustrates a problem that is **not** binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$, when the first draw selects a student. The probability is $\frac{6}{15}$, when the first draw selects a staff member.

The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

? Exercise 5.3.9

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

Answer

This is not binomial because the names are not replaced, which means the probability changes for each time a name is drawn. This violates the condition of independence.

References

1. "Access to electricity (% of population)," The World Bank, 2013. Available online at <http://data.worldbank.org/indicator/...first&sort=asc> (accessed May 15, 2015).
2. "Distance Education." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Distance_education (accessed May 15, 2013).
3. "NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).
4. Newport, Frank. "Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income," GALLUP® Economy, 2013. Available online at <http://www.gallup.com/poll/162368/am...-spending.aspx> (accessed May 15, 2013).
5. Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/N...eshman2011.pdf> (accessed May 15, 2013).
6. "The World FactBook," Central Intelligence Agency. Available online at <https://www.cia.gov/library/publicat...k/geos/af.html> (accessed May 15, 2013).
7. "What are the key statistics about pancreatic cancer?" American Cancer Society, 2013. Available online at www.cancer.org/cancer/pancrea...key-statistics (accessed May 15, 2013).

Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

There are a fixed number of trials, n .

There are only two possible outcomes, called "success" and, "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.

The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X = the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

Formula Review

- $X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p .
- X = the number of successes in n independent trials
- n = the number of independent trials
- X takes on the values $x = 0, 1, 2, 3, \dots, n$
- p = the probability of a success for any trial
- q = the probability of a failure for any trial
- $p + q = 1$

- $q = 1 - p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

Glossary

Binomial Experiment

a statistical experiment that satisfies the following three conditions:

1. There are a fixed number of trials, n .
2. There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

Bernoulli Trials

an experiment with the following characteristics:

1. There are only two possible outcomes called "success" and "failure" for each trial.
2. The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Binomial Probability Distribution

a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n , of independent trials.

"Independent" means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is

$$P(X = x) = \binom{n}{x} p^x q^{n-x}.$$

This page titled [5.3: Binomial Distribution](#) 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

6: Continuous Random Variables

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

[6.0: Introduction](#)

[6.1: Continuous Probability Functions](#)

[6.2: The Uniform Distribution](#)

This page titled [6: Continuous Random Variables](#) 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.

6.0: Introduction

CHAPTER OBJECTIVES

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

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.



Figure 6.0.1: The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Properties of Continuous Probability Distributions

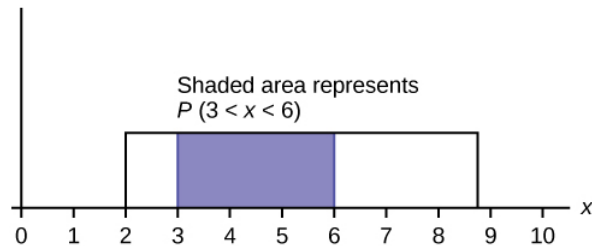
The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve. The curve is called the probability density function (abbreviated as pdf). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution. Area under the curve is given by a different function called the cumulative distribution function (abbreviated as cdf). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x -axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d . $P(c < x < d)$ is the area under the curve, above the x -axis, to the right of c and the left of d .
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x -axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this

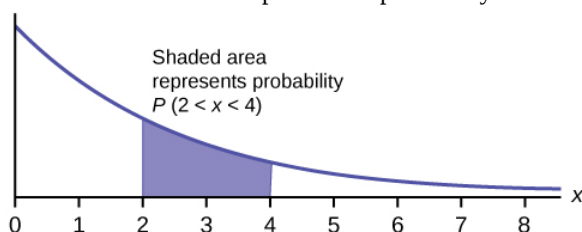
textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook. There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

In this chapter and the next, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.



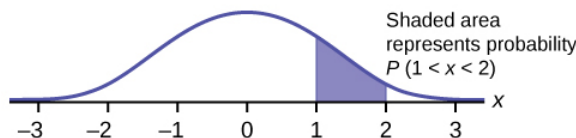
The uniform distribution

Figure 6.0.2 The graph shows a Uniform Distribution with the area between $x = 3$ and $x = 6$ shaded to represent the probability that the value of the random variable X is in the interval between



The exponential distribution

three and six. Figure 6.0.3 The graph shows an Exponential Distribution with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the



The normal distribution

interval between two and four. Figure 6.0.4 The graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

Glossary

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital; the notation is $X \sim \text{Exp}(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

This page titled 6.0: 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.

6.1: Continuous Probability Functions

We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x -axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. **For continuous probability distributions, PROBABILITY = AREA.**

✓ Example 6.1.1

Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$. x is a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.



Figure 6.1.1

$$f(x) = \frac{1}{20} \text{ for } 0 \leq x \leq 20. \quad (6.1.1)$$

The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the x -axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.

$$AREA = 20 \left(\frac{1}{20} \right) = 1 \quad (6.1.2)$$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x -axis where $0 < x < 2$.



Figure 6.1.2

$$AREA = (2 - 0) \left(\frac{1}{20} \right) = 0.1 \quad (6.1.3)$$

$$(2 - 0) = 2 = \text{base of a rectangle}$$

REMINDER: area of a rectangle = (base)(height).

The area corresponds to a probability. The probability that x is between zero and two is 0.1, which can be written mathematically as $P(0 < x < 2) = P(x < 2) = 0.1$.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x -axis where $4 < x < 15$.

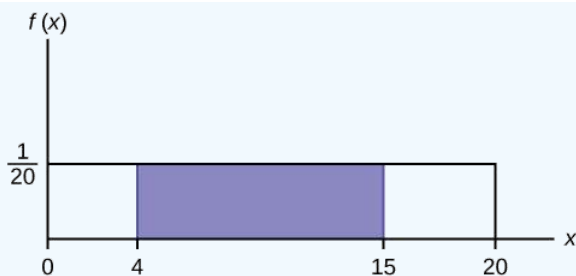


Figure 6.1.3

$$\text{AREA} = (15 - 4) \left(\frac{1}{20} \right) = 0.55$$

$$\text{AREA} = (15 - 4) \left(\frac{1}{20} \right) = 0.55$$

$(15 - 4) = 11 = \text{the base of a rectangle}$ $(15 - 4) = 11 = \text{the base of a rectangle}$

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

Suppose we want to find $P(x = 15)$. On an x-y graph, $x = 15$ is a vertical line. A vertical line has no width (or zero width). Therefore, $P(x = 15) = (\text{base})(\text{height}) = (0) \left(\frac{1}{20} \right) = 0$

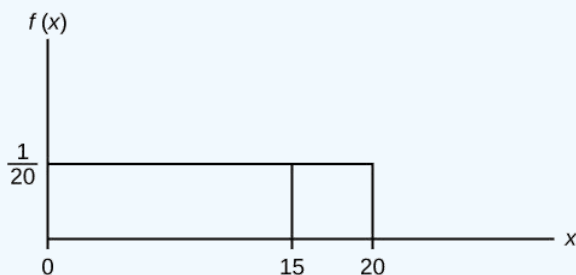


Figure 6.1.4

$P(X \leq x)$ (can be written as $P(X < x)$ for continuous distributions) is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right." We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 - P(X < x)$.

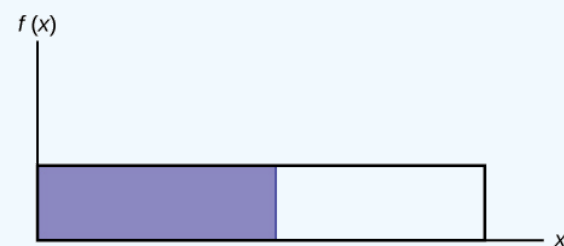


Figure 6.1.5

Label the graph with $f(x)$ and x . Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}$, $0 \leq x \leq 20$.

To calculate the probability that x is between two values, look at the following graph. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.

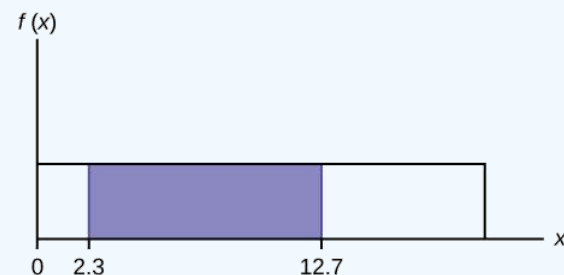


Figure 6.1.6

$$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3) \left(\frac{1}{20} \right) = 0.52 \quad (6.1.4)$$

? Exercise 6.1.1

Consider the function $f(x) = \frac{1}{8}$ for $0 \leq x \leq 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

Answer

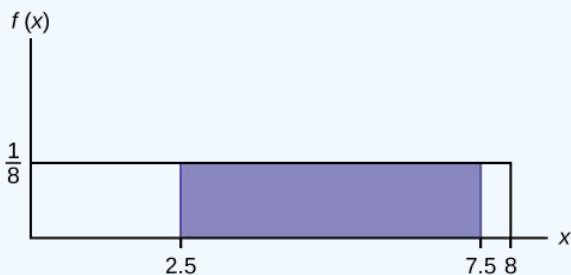


Figure 6.1.7

$$P(2.5 < x < 7.5) = 0.625$$

Summary

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points a and b is equal to $P(a < x < b)$. The cumulative distribution function (cdf) gives the probability as an area. If X is a continuous random variable, the probability density function (pdf), $f(x)$, is used to draw the graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(a < x < b)$.

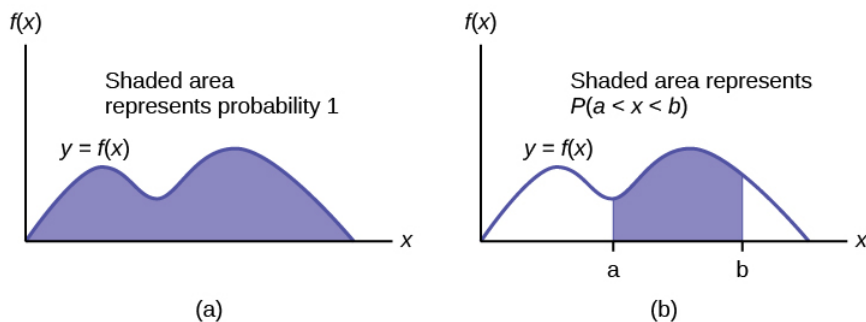


Figure 6.1.8

The cumulative distribution function (cdf) of X is defined by $P(X \leq x)$. It is a function of x that gives the probability that the random variable is less than or equal to x .

Formula Review

Probability density function (pdf) $f(x)$:

- $f(x) \geq 0$
- The total area under the curve $f(x)$ is one.

Cumulative distribution function (cdf): $P(X \leq x)$

This page titled [6.1: Continuous Probability Functions](#) 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.

6.2: The Uniform Distribution

The uniform distribution is a continuous probability distribution and is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data is inclusive or exclusive.

✓ Example 6.2.1

The data in Table 6.2.1 are 55 smiling times, in seconds, of an eight-week-old baby.

Table 6.2.1

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

The sample mean = 11.49 and the sample standard deviation = 6.23.

We will assume that the smiling times, in seconds, follow a uniform distribution between zero and 23 seconds, inclusive. This means that any smiling time from zero to and including 23 seconds is equally likely. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let X = length, in seconds, of an eight-week-old baby's smile.

The notation for the uniform distribution is

$X \sim U(a, b)$ where a = the lowest value of x and b = the highest value of x .

The probability density function is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.

For this example, $X \sim U(0, 23)$ and $f(x) = \frac{1}{23-0}$ for $0 \leq X \leq 23$.

Formulas for the theoretical mean and standard deviation are

$$\mu = \frac{a+b}{2}$$

and

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

For this problem, the theoretical mean and standard deviation are

$$\mu = \frac{0+23}{2} = 11.50 \text{ seconds}$$

and

$$\sigma = \frac{(23-0)^2}{12} = 6.64 \text{ seconds}.$$

Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation in this example.

? Exercise 6.2.1

The data that follow are the number of passengers on 35 different charter fishing boats. The sample mean = 7.9 and the sample standard deviation = 4.33. The data follow a uniform distribution where all values between and including zero and 14 are equally likely. State the values of a and b . Write the distribution in proper notation, and calculate the theoretical mean and standard deviation.

Table 6.2.2

1	12	4	10	4	14	11
7	11	4	13	2	4	6
3	10	0	12	6	9	10
5	13	4	10	14	12	11
6	10	11	0	11	13	2

Answer

a is zero; b is 14; $X \sim U(0, 14)$; $\mu = 7$ passengers; $\sigma = 4.04$ passengers

? Exercise 6.2.2

Refer to Example 6.3.1. What is the probability that a randomly chosen eight-week-old baby smiles between two and 18 seconds?

Answer

Find $P(2 < x < 18)$.

$$P(2 < x < 18) = (\text{base})(\text{height}) = (18 - 2) \left(\frac{1}{23} \right) = \left(\frac{16}{23} \right).$$

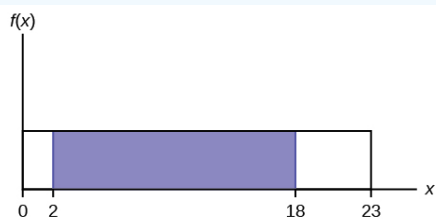


Figure 6.2.1 This graph shows a uniform distribution. The horizontal axis ranges from 0 to 15. The distribution is modeled by a rectangle extending from $x = 0$ to $x = 15$. A region from $x = 2$ to $x = 18$ is shaded inside the rectangle.

? Exercise 6.2.3

Find the 90th percentile for an eight-week-old baby's smiling time.

Answer

Ninety percent of the smiling times fall below the 90th percentile, k , so $P(x < k) = 0.90$

$$P(x < k) = 0.90 \quad (6.2.1)$$

$$(\text{base})(\text{height}) = 0.90 \quad (6.2.2)$$

$$(k - 0) \left(\frac{1}{23} \right) = 0.90 \quad (6.2.3)$$

$$k = (23)(0.90) = 20.7 \quad (6.2.4)$$

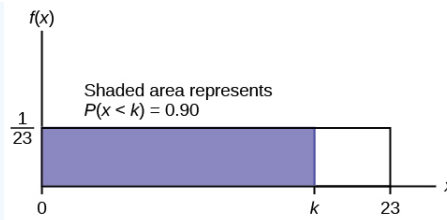


Figure 6.2.2 This shows the graph of the function $f(x) = 1/23$. A horizontal line ranges from the point $(0, 1/23)$ to the point $(23, 1/23)$. A vertical line extends from the x-axis to the end of the line at point $(k, 1/23)$ creating a rectangle. A region is shaded inside the rectangle from $x = 0$ to $x = k$. The shaded area represents $P(x < k) = 0.90$.

? Exercise 6.2.4

A distribution is given as $X \sim U(0, 20)$. What is $P(2 < x < 18)$? Find the 90th percentile.

Answer

$$P(2 < x < 18) = 0.8 ; 90^{\text{th}} \text{ percentile} = 18$$

? Exercise 6.2.5

The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.

What is the probability that a person waits fewer than 12.5 minutes?

Answer

Let X = the number of minutes a person must wait for a bus. $a = 0$ and $b = 15$. $X \sim U(0, 15)$. Write the probability density function. $f(x) = \frac{1}{15-0} = \frac{1}{15}$ for $0 \leq x \leq 15$.

Find $P(x < 12.5)$. Draw a graph.

$$P(x < k) = (\text{base})(\text{height}) = (12.5 - 0) \left(\frac{1}{15} \right) = 0.8333 \quad (6.2.5)$$

The probability a person waits less than 12.5 minutes is 0.8333.

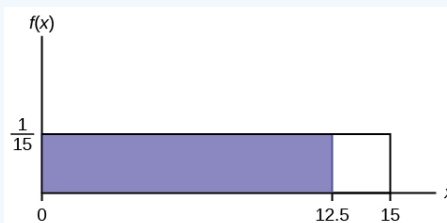


Figure 6.2.5. This shows the graph of the function $f(x) = 1/15$. A horizontal line ranges from the point $(0, 1/15)$ to the point $(15, 1/15)$. A vertical line extends from the x-axis to the end of the line at point $(12.5, 1/15)$ creating a rectangle. A region is shaded inside the rectangle from $x = 0$ to $x = 12.5$.

? Exercise 6.2.6

Refer to the previous exercise. Ninety percent of the time, the time a person must wait falls below what value?

Answer

Find the 90th percentile. Draw a graph. Let k = the 90th percentile.

$$P(x < k) = (\text{base})(\text{height}) = (k - 0) \left(\frac{1}{15} \right)$$

$$0.90 = (k) \left(\frac{1}{15} \right)$$

$$k = (0.90)(15) = 13.5$$

k is sometimes called a critical value.

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.

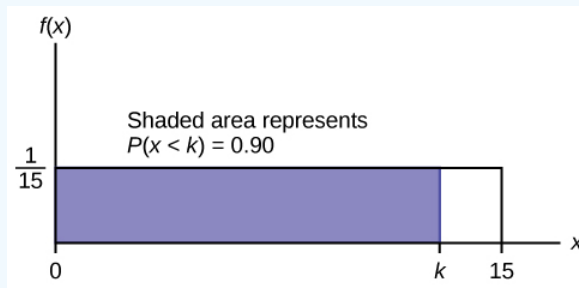


Figure 6.2.6.

? Exercise 6.2.7

The total duration of baseball games in the major league in the 2011 season is uniformly distributed between 447 hours and 521 hours inclusive.

- Find a and b and describe what they represent.
- Write the distribution.
- Find the mean and the standard deviation.
- What is the probability that the duration of games for a team for the 2011 season is between 480 and 500 hours?
- What is the 65th percentile for the duration of games for a team for the 2011 season?

Answer

- a is 447, and b is 521. a is the minimum duration of games for a team for the 2011 season, and b is the maximum duration of games for a team for the 2011 season.
- $X \sim U(447, 521)$
- $\mu = 484$, and $\sigma = 21.36$
- $P(480 < x < 500) = 0.2703$
- 65th percentile is 495.1 hours.

✓ Example 6.2.8

Suppose the time it takes a nine-year old to eat a donut is between 0.5 and 4 minutes, inclusive. Let X = the time, in minutes, it takes a nine-year old child to eat a donut. Then $X \sim U(0.5, 4)$.

The probability that a randomly selected nine-year old child eats a donut in at least two minutes is _____.

Solution

0.5714

✓ Example 6.2.9

Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and four hours. Let x = the time needed to fix a furnace. Then $x \sim U(1.5, 4)$.

- Find the probability that a randomly selected furnace repair requires more than two hours.
- Find the probability that a randomly selected furnace repair requires less than three hours.
- Find the 30th percentile of furnace repair times.

- d. The longest 25% of furnace repair times take at least how long? (In other words: find the minimum time for the longest 25% of repair times.) What percentile does this represent?
- e. Find the mean and standard deviation

Solution

a. To find $f(x)$: $f(x) = \frac{1}{4-1.5} = \frac{1}{2.5}$ so $f(x) = 0.4$

$$P(x > 2) = (\text{base})(\text{height}) = (4-2)(0.4) = 0.8$$

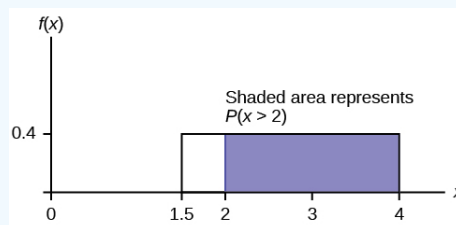


Figure 6.2.3. Uniform Distribution between 1.5 and four with shaded area between two and four representing the probability that the repair time x is greater than two

b. $P(x < 3) = (\text{base})(\text{height}) = (3-1.5)(0.4) = 0.6$

The graph of the rectangle showing the entire distribution would remain the same. However the graph should be shaded between $x = 1.5$ and $x = 3$. Note that the shaded area starts at $x = 1.5$ rather than at $x = 0$; since $X \sim U(1.5, 4)$, x can not be less than 1.5.

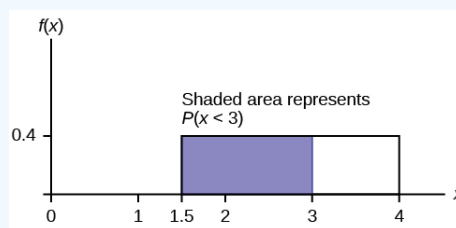


Figure 6.2.4. Uniform Distribution between 1.5 and four with shaded area between 1.5 and three representing the probability that the repair time x is less than three

c.

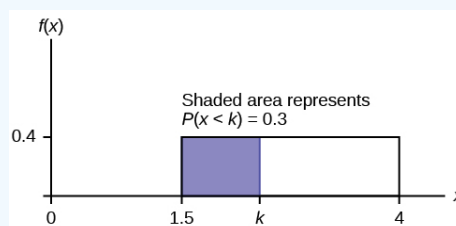


Figure 6.2.5. Uniform Distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30% of repair times.

$$P(x < k) = 0.30$$

$$P(x < k) = (\text{base})(\text{height}) = (k-1.5)(0.4)$$

$$0.3 = (k-1.5)(0.4)$$

Solve for k by dividing both sides by 0.4 and then adding 1.5 to both sides:

$$0.75 = k - 1.5$$

$$k = 2.25$$

The 30th percentile of repair times is 2.25 hours. 30% of repair times are 2.25 hours or less.

d.

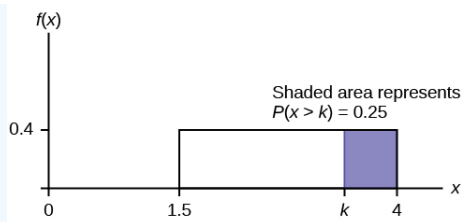


Figure 6.2.6. Uniform Distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25% of repair times.

$$P(x > k) = 0.25$$

$$P(x > k) = (\text{base})(\text{height}) = (4 - k)(0.4)$$

$$0.25 = (4 - k)(0.4)$$

Solve for k by dividing both sides by 0.4 and then subtracting four from both sides:

$$0.625 = 4 - k,$$

$$-3.375 = -k,$$

$$k = 3.375$$

The longest 25% of furnace repairs take at least 3.375 hours (3.375 hours or longer).

Note: Since 25% of repair times are 3.375 hours or longer, that means that 75% of repair times are 3.375 hours or less. 3.375 hours is the **75th percentile** of furnace repair times.

$$\text{e. } \mu = \frac{a+b}{2} \text{ and } \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

$$\mu = \frac{1.5+4}{2} = 2.75 \text{ hours and } \sigma = \sqrt{\frac{(4-1.5)^2}{12}} = 0.7217 \text{ hours}$$

? Exercise 6.2.10

The amount of time a service technician needs to change the oil in a car is uniformly distributed between 11 and 21 minutes. Let X = the time needed to change the oil on a car.

- Write the random variable X in words. X = _____.
- Write the distribution.
- Graph the distribution.
- Find $P(x > 19)$.
- Find the 50th percentile.

Answer

- Let X = the time needed to change the oil in a car.
- $X \sim U(11, 21)$.
-  Your graph should show a uniform distribution. The horizontal axis can range from 405 to 525. The distribution is modeled by a rectangle extending from $x = 447$ to $x = 521$.
- $P(x > 19) = 0.2$
- the 50th percentile is 16 minutes.

Use the following information to answer the next four questions. The data that follow are the square footage (in 1,000 feet squared) of 28 homes.

1.5	2.4	3.6	2.6	1.6	2.4	2.0
3.5	2.5	1.8	2.4	2.5	3.5	4.0
2.6	1.6	2.2	1.8	3.8	2.5	1.5
2.8	1.8	4.5	1.9	1.9	3.1	1.6

The sample mean = 2.50 and the sample standard deviation = 0.8302.

The distribution can be modeled as $X \sim U(1.5, 4.5)$.

? Exercise 6.2.11

In this distribution, outcomes are equally likely. What does this mean?

Answer

It means that the value of x is just as likely to be any number between 1.5 and 4.5.

? Exercise 6.2.12

What are the constraints for the values of x ?

Answer

$$1.5 \leq x \leq 4.5$$

? Exercise 6.2.13

What is $P(2 < x < 3)$?

Answer

0.3333

? Exercise 6.2.14

What is $P(x = 1.5)$?

Answer

zero

Use the following information to answer the next six exercises.

The age of cars in the staff parking lot of a suburban college is uniformly distributed from six months (0.5 years) to 9.5 years.

? Exercise 6.2.15

In words, define the random variable X .

Answer

X = The age (in years) of cars in the staff parking lot

? Exercise 6.2.16

The interval of values for x is _____.

Answer

0.5 to 9.5

? Exercise 6.2.17

Write the probability density function.

Answer

$$f(x) = \frac{1}{9} \text{ where } x \text{ is between } 0.5 \text{ and } 9.5, \text{ inclusive.}$$

? Exercise 6.2.18

Find the average age of the cars in the lot.

Answer

$$\mu = 5$$

? Exercise 6.2.19

Considering only the cars less than 7.5 years old, find the probability. $P(4 < x < 7.5) = \underline{\hspace{2cm}}$

Answer

$$\frac{3.5}{7}$$

? Exercise 6.2.20

Find the third quartile of ages of cars in the lot. This means you will have to find the value such that $\frac{3}{4}$, or 75%, of the cars are at most (less than or equal to) that age.

Answer

$$7.25$$

Review

If X has a uniform distribution where $a < x < b$ or $a \leq x \leq b$, then X takes on values between a and b (may include a and b). All values x are equally likely. We write $X \sim U(a, b)$. The mean of X is $\mu = \frac{a+b}{2}$. The standard deviation of X is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function of X is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$. The cumulative distribution function of X is $P(X \leq x) = \frac{x-a}{b-a}$. X is continuous.

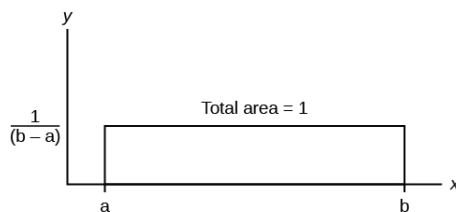


Figure 6.2.7.

The probability $P(c < X < d)$ may be found by computing the area under $f(x)$, between c and d . Since the corresponding area is a rectangle, the area may be found simply by multiplying the width and the height.

Formula Review

X = a real number between a and b (in some instances, X can take on the values a and b). a = smallest X ; b = largest X

$$X \sim U(a, b)$$

The mean is $\mu = \frac{a+b}{2}$

The standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Probability density function: $f(x) = \frac{1}{b-a}$ for $a \leq X \leq b$

Area to the Left of x : $P(X < x) = (x-a) \left(\frac{1}{b-a} \right)$

Area to the Right of x : $P(X > x) = (b-x) \left(\frac{1}{b-a} \right)$

Area Between c and d : $P(c < x < d) = (\text{base})(\text{height}) = (d-c) \left(\frac{1}{b-a} \right)$

Uniform: $X \sim U(a, b)$ where $a < x < b$

- pdf: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- cdf: $P(X \leq x) = \frac{x-a}{b-a}$
- mean $\mu = \frac{a+b}{2}$
- standard deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- $P(c < X < d) = (d-c) \left(\frac{1}{b-a} \right)$

References

McDougall, John A. The McDougall Program for Maximum Weight Loss. Plume, 1995.

Glossary

Conditional Probability

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

This page titled [6.2: The Uniform Distribution](#) 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: The Normal Distribution

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ).

[7.0: Prelude to The Normal Distribution](#)

[7.1: The Standard Normal Distribution](#)

[7.2: Using the Normal Distribution](#)

This page titled [7: The Normal Distribution](#) 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.

7.0: Prelude to The Normal Distribution

Learning Objectives

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

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.



Figure 7.0.1 If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{1}{2}\right) \cdot \left(\frac{x - \mu}{\sigma}\right)^2} \quad (7.0.1)$$

The probability density function is a rather complicated function. **Do not memorize it.** It is not necessary.

The cumulative distribution function is $P(X < x)$. It is calculated either by a calculator or a computer, or it is looked up in a table. Technology has made the tables virtually obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions.

NORMAL: $X \sim N(\mu, \sigma)$

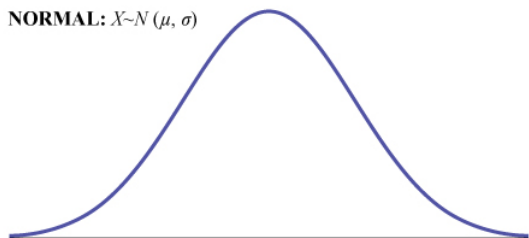


Figure 7.0.2 The standard normal distribution

The curve is symmetrical about a vertical line drawn through the mean, μ . In theory, the mean is the same as the median, because the graph is symmetric about μ . As the notation indicates, the normal distribution depends only on the mean and the standard

deviation. Since the area under the curve must equal one, a change in the standard deviation, σ , causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ . A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

COLLABORATIVE CLASSROOM ACTIVITY

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the x -axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

- $X \sim N(\mu, \sigma)$
- μ = the mean σ = the standard deviation

Glossary

Normal Distribution

a continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (7.0.2)$$

, 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 the **standard normal distribution**.

This page titled [7.0: Prelude to The Normal Distribution](#) 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.

7.1: The Standard Normal Distribution

Z-Scores

The standard normal distribution is a normal distribution of standardized values called *z-scores*. A *z-score* is measured in units of the standard deviation.

Definition: Z-Score

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z-score* is:

$$z = \frac{x - \mu}{\sigma} \quad (7.1.1)$$

The *z-score* tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ . Values of x that are larger than the mean have positive *z-scores*, and values of x that are smaller than the mean have negative *z-scores*. If x equals the mean, then x has a *z-score* of zero. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$$\begin{aligned} x &= \mu + (z)(\sigma) \\ &= 5 + (3)(2) = 11 \end{aligned}$$

The *z-score* is three.

Since the mean for the standard normal distribution is zero and the standard deviation is one, then the transformation in Equation 7.1.1 produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ .

*A **z-score** is measured in units of the standard deviation.*

✓ Example 7.1.1

Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then (via Equation 7.1.1):

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **two** standard deviations (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Notice that: $5 + (2)(6) = 17$ (The pattern is $\mu + z\sigma = x$)

Now suppose $x = 1$. Then:

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$$

(rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that: $5 + (-0.67)(6)$ is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ . Or, when z is positive, x is greater than μ , and when z is negative x is less than μ .

? Exercise 7.1.1

What is the *z-score* of x , when $x = 1$ and $X \sim N(12, 3)$?

Answer

$$z = \frac{1 - 12}{3} \approx -3.67$$

? Exercise 7.1.2

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

- Suppose a person **lost** ten pounds in a month. The z -score when $x = 10$ pounds is $z = 2.5$ (verify). This z -score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
- Suppose a person **gained** three pounds (a negative weight loss). Then $z =$ _____. This z -score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Answer

- This z -score tells you that $x = 10$ is 2.5 standard deviations to the right of the mean five.
- Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z ?

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2$$

where $\mu = 2$ and $\sigma = 1$.

The z -score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of their own) standard deviations to the right of their respective means.

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

? Exercise 7.1.3

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores ten points in a game. The z -score when $x = 10$ is -1.5 . This score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean value of _____ (What is the mean?).

Answer

1.5, left, 16

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ , then the *Empirical Rule* says the following:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z -scores for $+1\sigma$ and -1σ are $+1$ and -1 , respectively.
- The z -scores for $+2\sigma$ and -2σ are $+2$ and -2 , respectively.
- The z -scores for $+3\sigma$ and -3σ are $+3$ and -3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.

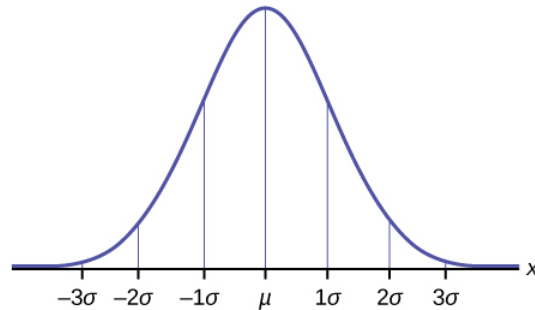


Figure 7.1.1

? Exercise 7.1.4

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

- Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z -score when $x = 168$ cm is $z =$ _____. This z -score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
- Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z -score of $z = 1.27$. What is the male's height? The z -score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answer

- 0.32, 0.32, left, 170
- 177.98, 1.27, right

? Exercise 7.1.5

Use the information in Exercise 7.1.4 to answer the following questions.

- Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z -score when $x = 176$ cm is $z =$ _____. This z -score tells you that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
- Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z -score of $z = -2$. What is the male's height? The z -score ($z = -2$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answer

- Solve the equation $z = \frac{x - \mu}{\sigma}$ for z . $x = \mu + (z)(\sigma)$

$z = \frac{176 - 170}{6.28}$, This z -score tells you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm.

- Solve the equation $z = \frac{x - \mu}{\sigma}$ for z . $x = \mu + (z)(\sigma)$

$X = 157.44$ cm, The z -score ($z = -2$) tells you that the male's height is two standard deviations to the left of the mean.

? Example 7.1.6

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the z -scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z -score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Answer

- The z -score (Equation 7.1.1) for $x = 160.58$ is $z = -1.5$.
- The z -score for $y = 162.85$ is $z = -1.5$.

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

? Exercise 7.1.7

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z -scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z -score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Answer

The z -score (Equation 7.1.1) for $x_1 = 325$ is $z_1 = -1.15$.

The z -score (Equation 7.1.1) for $x_2 = 366.21$ is $z_2 = -1.14$.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z -scores, Student 2 had the better score.

✓ Example 7.1.8

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie within one standard deviation of the mean. Therefore, about 68% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation from the mean 50. The z -scores are -1 and $+1$ for 44 and 56, respectively.
- About 95% of the x values lie within two standard deviations of the mean. Therefore, about 95% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within two standard deviations from the mean 50. The z -scores are -2 and $+2$ for 38 and 62, respectively.
- About 99.7% of the x values lie within three standard deviations of the mean. Therefore, about 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ from the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The z -scores are -3 and $+3$ for 32 and 68, respectively.

? Exercise 7.1.8

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Answer

between 20 and 30.

? Exercise 7.1.9

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

- About 68% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.
- About 95% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.
- About 99.7% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.

Answer

- About 68% of the values lie between 166.02 and 178.7. The z -scores are -1 and 1 .
- About 95% of the values lie between 159.68 and 185.04. The z -scores are -2 and 2 .
- About 99.7% of the values lie between 153.34 and 191.38. The z -scores are -3 and 3 .

? Exercise 7.1.10

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

- About 68% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.
- About 95% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.
- About 99.7% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.

Answer for part a

About 68% of the values lie between the values 41 and 63. The z -scores are -1 and 1 , respectively.

Answer for part b

About 95% of the values lie between the values 30 and 74. The z -scores are -2 and 2 , respectively.

Answer for part c

About 99.7% of the values lie between the values 19 and 85. The z -scores are -3 and 3 , respectively.

Summary

A z -score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z -scores is zero and the standard deviation is one. If y is the z -score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ .

Formula Review

$$Z \sim N(0, 1)$$

$$z = a \text{ standardized value (z-score)}$$

$$\text{mean} = 0; \text{ standard deviation} = 1$$

To find the K^{th} percentile of X when the z -scores is known:

$$k = \mu + (z)\sigma$$

$$z\text{-score: } z = \frac{x - \mu}{\sigma}$$

Z = the random variable for z-scores

$Z \sim N(0, 1)$

Glossary

Standard Normal Distribution

a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $(Z \sim N(0, 1))$.

z-score

the linear transformation of the form $z = \frac{x - \mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ , the result is called the z-score of x . The z-score allows us to compare data that are normally distributed but scaled differently.

References

1. "Blood Pressure of Males and Females." StatCrunch, 2013. Available online at <http://www.statcrunch.com/5.0/viewre...reportid=11960> (accessed May 14, 2013).
2. "The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).
3. "2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at media.collegeboard.com/digita...Group-2012.pdf (accessed May 14, 2013).
4. "Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009." National Center for Education Statistics. Available online at nces.ed.gov/programs/digest/d...s/dt09_147.asp (accessed May 14, 2013).
5. Data from the *San Jose Mercury News*.
6. Data from *The World Almanac and Book of Facts*.
7. "List of stadiums by capacity." Wikipedia. Available online at en.Wikipedia.org/wiki/List_o...ms_by_capacity (accessed May 14, 2013).
8. Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

This page titled [7.1: The Standard Normal Distribution](#) 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.

- [6.2: The Standard Normal Distribution](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

7.2: Using the Normal Distribution

The shaded area in the following graph indicates the area to the left of x . This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.

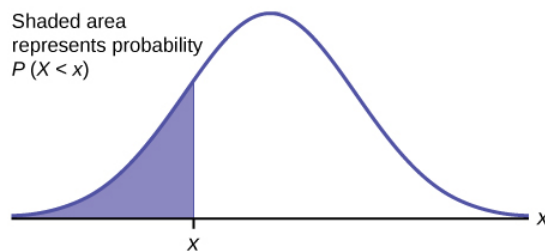


Figure 7.2.1.

The area to the right is then $P(X > x) = 1 - P(X < x)$. Remember, $P(X < x)$ = **area to the left** of the vertical line through x . $P(X > x) = 1 - P(X < x)$ = **area to the right** of the vertical line through x . $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using either technology or tables. There are instructions embedded in Exercise 7.2.2 for TI-83+ and TI-84 calculators. To calculate the probability using a table, use the guidelines indicated in the lecture notes for the course.

✓ Example 7.2.1

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$

? Exercise 7.2.1

If the area to the left of x is 0.012, then what is the area to the right?

Answer

$$1 - 0.012 = 0.988$$

? Example 7.2.2

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

- Find the probability that a randomly selected student scored more than 65 on the exam.
- Find the probability that a randomly selected student scored less than 85.
- Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).
- Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Answer for part a

- Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$$P(x > 65) = 0.3446$$

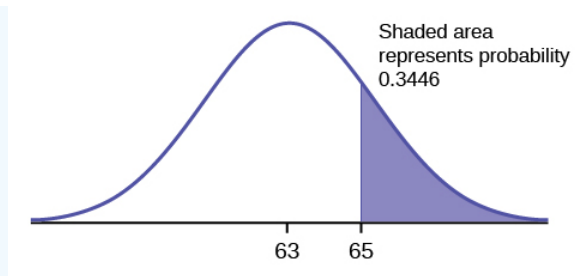


Figure 7.2.2.

The probability that any student selected at random scores more than 65 is 0.3446.

📌 USING THE TI-83, 83+, 84, 84+ CALCULATOR

Go into `2nd DISTR` .

After pressing `2nd DISTR` , press `2:normalcdf` .

The syntax for the instructions are as follows:

`normalcdf(lower value, upper value, mean, standard deviation)` For this problem: `normalcdf(65,1E99,63,5) = 0.3446`. You get `1E99` ($= 10^{99}$) by pressing `1` , the `EE` key (a 2nd key) and then `99` . Or, you can enter `10^ 99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99} . In some instances, the lower number of the area might be $-1E99$ ($= -10^{99}$). The number -10^{99} is way out in the left tail of the normal curve.

📌 Historical Note

The TI probability program calculates a z -score and then the probability from the z -score. Before technology, the z -score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z -score was used. You calculate the z -score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65-63}{5} = 0.4$$

Area to the left is 0.6554.

$$P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$$

📌 USING THE TI-83, 83+, 84, 84+ CALCULATOR

Find the percentile for a student scoring 65:

*Press `2nd Distr`

*Press `2:normalcdf` (

*Enter lower bound, upper bound, mean, standard deviation followed by)

*Press `ENTER` .

For this Example, the steps are

`2nd Distr`

`2:normalcdf (65,1 2nd EE 99,63,5) ENTER`

The probability that a selected student scored more than 65 is 0.3446.

To find the probability that a selected student scored *more than* 65, subtract the percentile from 1.

Answer for part b

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, find $P(x < 85) = 1$.

$\text{normalcdf}(0, 85, 63, 5) = 1$ (rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Answer for part c

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x -axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. The variable k is located on the x -axis. $P(x < k)$ is the area to the left of k . The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k , and ten percent are the same or higher. The variable k is often called a critical value.

$k = 69.4$

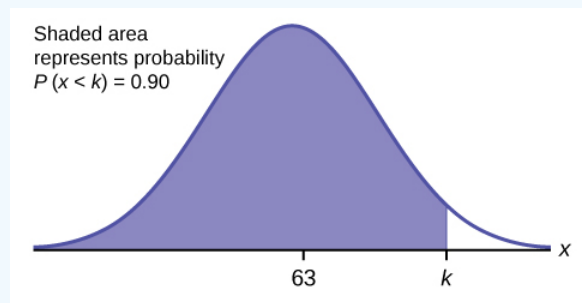


Figure 7.2.3.

The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. To get this answer on the calculator, follow this step:

`invNorm` in `2nd DISTR` . `invNorm`(area to the left, mean, standard deviation)

For this problem, $\text{invNorm}(0.90, 63, 5) = 69.4$

Answer for part d

d. Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.6 and 30% fall at or above.

$$\text{invNorm}(0.70, 63, 5) = 65.6$$

? Exercise 7.2.3

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a randomly selected golfer scored less than 65.

Answer

$$\text{normalcdf}(-10^{99}, 65, 68, 3) = 0.1587$$

? Exercise 7.2.4

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is

used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

- Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
- Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Answer

a. Let X = the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

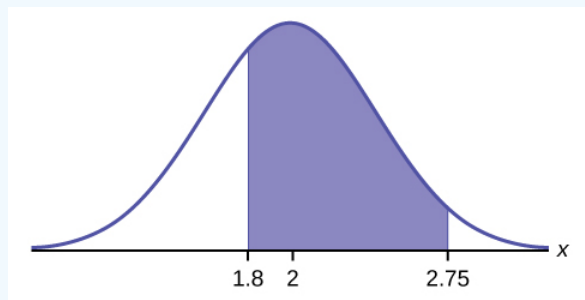


Figure 7.2.4.

$$\text{normalcdf}(1.8, 2.75, 2, 0.5) = 0.5886$$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile, k** , where $P(x < k) = 0.25$.

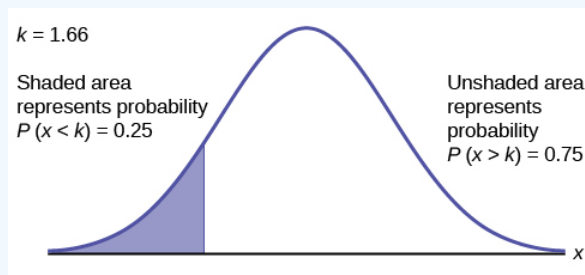


Figure 7.2.5.

$$\text{invNorm}(0.25, 2, 0.5) = 1.66$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

? Exercise 7.2.5

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Answer

$$\text{normalcdf}(66, 70, 68, 3) = 0.4950$$

? Exercise 7.2.6

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

- Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.
- Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.
- Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Answer

- $\text{normalcdf}(23, 64.7, 36.9, 13.9) = 0.8186$
- $\text{normalcdf}(-10^{99}, 50.8, 36.9, 13.9) = 0.8413$
- $\text{invNorm}(0.80, 36.9, 13.9) = 48.6$

The 80th percentile is 48.6 years.

80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

Use the information in previous exercise to answer the next question.

? Exercise 7.2.7

- Find the 30th percentile, and interpret it in a complete sentence.
- What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old?

Answer

- Let X = a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$

To find the 30th percentile, find k such that $P(x < k) = 0.30$.

$\text{invNorm}(0.30, 36.9, 13.9) = 29.6$ years

Thirty percent of smartphone users 13 to 55+ are at most 29.6 years and 70% are at least 29.6 years.

- Find $P(x < 27)$

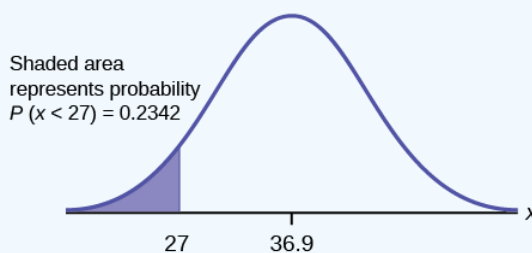


Figure 7.2.6.

$$\text{normalcdf}(0, 27, 36.9, 13.9) = 0.2342$$

(Note that $\text{normalcdf}(-10^{99}, 27, 36.9, 13.9) = 0.2382$. The two answers differ only by 0.0040.)

? Exercise 7.2.8

In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years respectively. Using this information, answer the following questions (round answers to one decimal place).

- Calculate the interquartile range (*IQR*).
- Forty percent of the ages that range from 13 to 55+ are at least what age?

Answer

a.

$$IQR = Q_3 - Q_1$$

Calculate $Q_3 = 75^{\text{th}}$ percentile and $Q_1 = 25^{\text{th}}$ percentile.

$$\text{invNorm}(0.75, 36.9, 13.9) = Q_3 = 46.2754$$

$$\text{invNorm}(0.25, 36.9, 13.9) = Q_1 = 27.5246$$

$$IQR = Q_3 - Q_1 = 18.7508$$

b. Find k where $P(x > k) = 0.40$ ("At least" translates to "greater than or equal to.")

$0.40 =$ the area to the right.

Area to the left $= 1 - 0.40 = 0.60$

The area to the left of $k = 0.60$.

$$\text{invNorm}(0.60, 36.9, 13.9) = 40.4215$$

$k = 40.42$.

Forty percent of the smartphone users from 13 to 55+ are at least 40.4 years.

? Exercise 7.2.9

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean $\mu = 81$ points and standard deviation $\sigma = 15$ points.

- Calculate the first- and third-quartile scores for this exam.
- The middle 50% of the exam scores are between what two values?

Answer

$$\text{a. } Q_1 = 25^{\text{th}} \text{ percentile} = \text{invNorm}(0.25, 81, 15) = 70.9$$

$$Q_3 = 75^{\text{th}} \text{ percentile} = \text{invNorm}(0.75, 81, 15) = 91.1$$

b. The middle 50% of the scores are between 70.9 and 91.1.

? Example 7.2.10

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

- Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
- The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.
- Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Answer

$$\text{a. } \text{normalcdf}(6, 10^{99}, 5.85, 0.24) = 0.2660$$

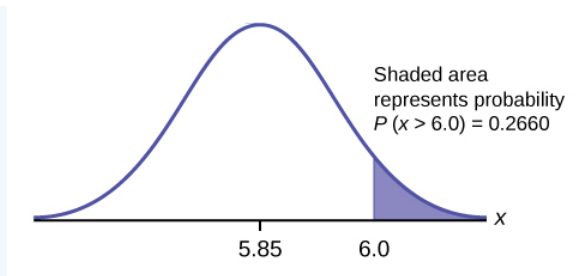


Figure 7.2.7.

Answer

b. $1 - 0.20 = 0.80$

The tails of the graph of the normal distribution each have an area of 0.40.

Find k_1 , the 40th percentile, and k_2 , the 60th percentile ($0.40 + 0.20 = 0.60$).

$k_1 = \text{invNorm}(0.40, 5.85, 0.24) = 5.79\text{m}$

$k_2 = \text{invNorm}(0.60, 5.85, 0.24) = 5.91\text{m}$

c. 6.16: Ninety percent of the diameter of the mandarin oranges is at most 6.15 cm.

? Exercise 7.2.11

Using the information from the previous exercise, answer the following:

- The middle 45% of mandarin oranges from this farm are between _____ and _____.
- Find the 16th percentile and interpret it in a complete sentence.

Answer for part a

The middle area = 0.40, so each tail has an area of 0.30.

$1 - 0.40 = 0.60$

The tails of the graph of the normal distribution each have an area of 0.30.

Find k_1 , the 30th percentile and k_2 , the 70th percentile ($0.40 + 0.30 = 0.70$).

$k_1 = \text{invNorm}(0.30, 5.85, 0.24) = 5.72\text{m}$

$k_2 = \text{invNorm}(0.70, 5.85, 0.24) = 5.98\text{m}$

Answer for part b

$\text{normalcdf}(5, 10^{99}, 5.85, 0.24) = 0.9998$

? Exercise 7.2.12

How would you represent the area to the left of one in a probability statement?

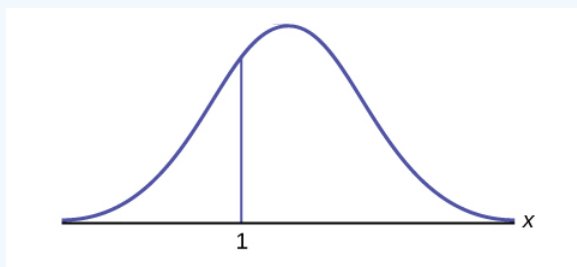


Figure 7.2.8.

Answer

$$P(x < 1)$$

? Exercise 7.2.13

Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

Answer

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

? Exercise 7.2.14

How would you represent the area to the right of three in a probability statement?

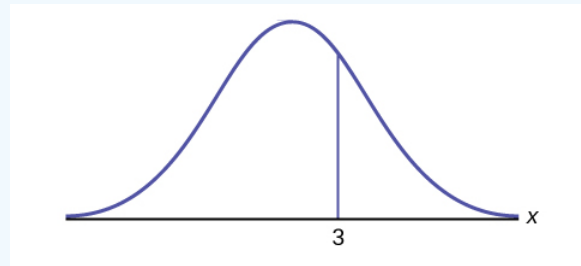


Figure 7.2.9.

Answer

$$1 - P(x < 3) \text{ or } P(x > 3)$$

? Exercise 7.2.15

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x ?

Answer

$$1 - 0.543 = 0.457$$

Use the following information to answer the next two exercises:

$$X \sim N(54, 8)$$

? Exercise 7.2.16

Find the probability that $x < 30$.

Answer

$$0.0013$$

? Exercise 7.2.17

Find the 60th percentile.

Answer

$$56.03$$

? Exercise 7.2.18

$$X \sim N(-3, 4)$$

Find the probability that x is between one and four.

Answer

0.1186

? Exercise 7.2.19

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

- a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

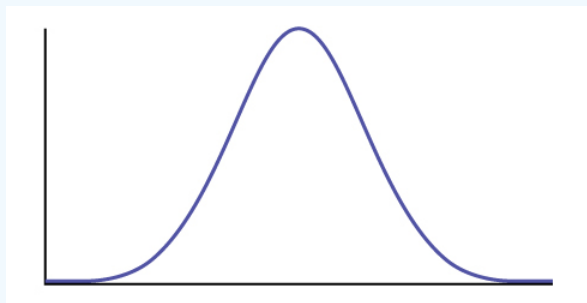


Figure 7.2.10.

$$P(0 < x < \text{_____}) = \text{_____} \text{ (Use zero for the minimum value of } x\text{.)}$$

Answer

- a. Post your sketch on the discussion boards for review.
b. 3, 0.1979

? Exercise 7.2.20

Find the 70th percentile of the distribution for the time a CD player lasts.

- a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.

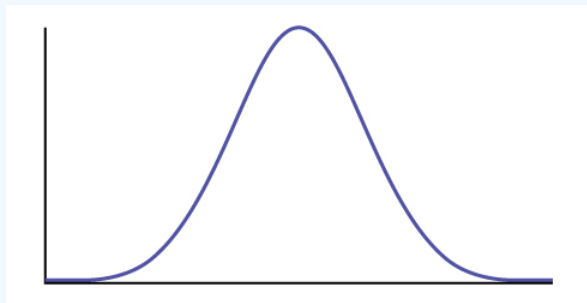


Figure 7.2.11.

$$P(x < k) = \text{_____} \text{ Therefore, } k = \text{_____}$$

Answer

- a. Post your sketch on the discussion boards for review.
b. 0.70, 4.78 years

References

1. "Naegle's rule." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Naegle's_rule (accessed May 14, 2013).
2. "403: NUMMI." Chicago Public Media & Ira Glass, 2013. Available online at www.thisamericanlife.org/radiosode/403/nummi (accessed May 14, 2013).
3. "Scratch-Off Lottery Ticket Playing Tips." WinAtTheLottery.com, 2013. Available online at www.winatthelottery.com/publication/partment40.cfm (accessed May 14, 2013).
4. "Smart Phone Users, By The Numbers." Visual.ly, 2013. Available online at <http://visual.ly/smart-phone-users-numbers> (accessed May 14, 2013).
5. "Facebook Statistics." Statistics Brain. Available online at <http://www.statisticbrain.com/facebook-statistics/> (accessed May 14, 2013).

Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ . A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

- Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.
- Standard Normal Distribution: $Z \sim N(0, 1)$.
- Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)
- Calculator function for the k^{th} percentile: $k = \text{invNorm}$ (area to the left of k , mean, standard deviation)

This page titled [7.2: Using the Normal Distribution](#) 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.

- [6.3: Using the Normal Distribution](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

CHAPTER OVERVIEW

8: The Central Limit Theorem

In this chapter, you will study means and the **central limit theorem**, which is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ , and a known standard deviation, σ . The first alternative says that if we collect samples of size n with a "large enough n ," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

[8.0: Prelude to the Central Limit Theorem](#)

[8.1: The Central Limit Theorem for Sample Means \(Averages\)](#)

[8.2: The Central Limit Theorem for Sums](#)

[8.3: Using the Central Limit Theorem](#)

Template:ContribOpenStax

This page titled [8: The Central Limit Theorem](#) 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.

8.0: Prelude to the Central Limit Theorem

Learning Objectives

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

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.
- Apply and interpret the central limit theorem for sums.

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**. The **central limit theorem** (clt for short) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ , and a known standard deviation, σ . The first alternative says that if we collect samples of size n with a "large enough n ," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means and the sums tend to follow the normal distribution.

The size of the sample, n , that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement.**



Figure 8.0.1 If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

COLLABORATIVE CLASSROOM ACTIVITY

Suppose eight of you roll one fair die ten times, seven of you roll two fair dice ten times, nine of you roll five fair dice ten times, and 11 of you roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample **mean** of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5} = 3.4$. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

Your instructor will pass out the dice to several people. Roll your dice ten times. For each roll, record the faces, and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for one die, one graph for two dice, one graph for five dice, and one graph for ten dice. Since the "mean" when you roll one die is just the face on the die, what distribution do these **means** appear to be representing?

- **Draw the graph for the means using two dice.** Do the sample means show any kind of pattern?
- **Draw the graph for the means using five dice.** Do you see any pattern emerging?
- **Finally, draw the graph for the means using ten dice.** Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from one to two to five to ten, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the central limit theorem (clt). The central limit theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

Glossary

Sampling Distribution

Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

This page titled [8.0: Prelude to the Central Limit Theorem](#) 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.

8.1: The Central Limit Theorem for Sample Means (Averages)

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:

- μ_x = the mean of X
- σ_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

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right). \quad (8.1.1)$$

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.

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)} \quad (8.1.2)$$

- 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

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

Exercise 8.1.1

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.

- Find the probability that the sample mean is between 85 and 92.
- Find the value that is two standard deviations above the expected value, 90, of the sample mean.

Answer

- 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.

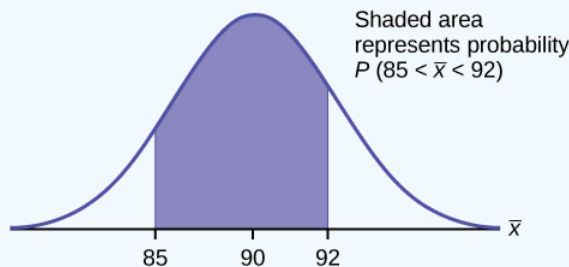


Figure 8.1.1.

`normalcdf` (lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ , $\frac{\sigma}{\sqrt{n}}$)

$$\text{normalcdf} \left(85, 92, 90, \frac{15}{\sqrt{25}} \right) = 0.6997$$

b. To find the value that is two standard deviations above the expected value 90, use the formula:

$$\begin{aligned} \text{value} &= \mu_x + (\# \text{ of TSDEVs}) \left(\frac{\sigma_x}{\sqrt{n}} \right) \\ &= 90 + 2 \left(\frac{15}{\sqrt{25}} \right) = 96 \end{aligned}$$

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 8.1.2

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

$$P(42 < \bar{x} < 50) = \left(42, 50, 45, \frac{8}{\sqrt{30}} \right) = 0.9797$$

? Exercise 8.1.3

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 $X \sim N(\text{_____, ____})$ by the central limit theorem for means.

$$\mu_x = 2, \sigma_x = 0.5, n = 50, \text{ and } X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$$

Find $P(1.8 < \bar{x} < 2.3)$. Draw a graph.

$$P(1.8 < \bar{x} < 2.3) = 0.9977$$

$$\text{normalcdf}\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

? Exercise 8.1.4

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

$$x = \text{invNorm}\left(\text{area to the left of } x, \text{mean}, \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}\right)$$

where:

- x = the x value that corresponds to the k^{th} percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

? Exercise 8.1.5

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 $n = 100$.

- What are the mean and standard deviation for the sample mean ages of tablet users?
- What does the distribution look like?
- Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
- Find the 95th percentile for the sample mean age (to one decimal place).

Answer

- Since the sample mean tends to target the population mean, we have $\mu_x = \mu = 34$. The sample standard deviation is given by:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

- The central limit theorem states that for large sample sizes (n), the sampling distribution will be approximately normal.
- The probability that the sample mean age is more than 30 is given by:

$$P(X > 30) = \text{normalcdf}(30, E99, 34, 1.5) = 0.9962$$

- Let k = the 95th percentile.

$$k = \text{invNorm}\left(0.95, 34, \frac{15}{\sqrt{100}}\right) = 36.5$$

? Exercise 8.1.6

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.

$$P(29 < \bar{x} < 35) = \text{normalcdf}\left(29, 35, 28, \frac{4.8}{\sqrt{100}}\right) = 0.0186 \quad (8.1.3)$$

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.

? Exercise 8.1.7

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.

- What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
- What is the standard error of the mean?
- Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
- Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Answer

- $\mu = \mu = 8.2, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$
- This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.
- Let k = the 90th percentile
 $k = \text{invNorm}\left(0.90, 8.2, \frac{1}{\sqrt{60}}\right) = 8.37$ This value indicates that 90 percent of the average app engagement time for table users is less than 8.37 minutes.
- $P(8 < \bar{x} < 8.5) = \text{normalcdf}\left(8, 8.5, 8.2, \frac{1}{\sqrt{60}}\right) = 0.9293$

? Exercise 8.1.8

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 $\bar{X} : \sigma_x$
- 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 [8.1: The Central Limit Theorem for Sample Means \(Averages\)](#) 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.

8.2: The Central Limit Theorem for Sums

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution) and suppose:

- μ_x = the mean of X
- σ_x = the standard deviation of X

If you draw random samples of size n , then as n increases, the random variable $\sum X$ consisting of sums tends to be normally distributed and

$$\sum X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x)). \quad (8.2.1)$$

The central limit theorem for sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample size increases. The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.

The random variable $\sum X$ has the following z-score associated with it:

- $\sum x$ is one sum.
- $z = \frac{\sum x - (n)(\mu_x)}{(\sqrt{n})(\sigma_x)}$
 - $(n)(\mu_x)$ = the mean of $\sum X$
 - $(\sqrt{n})(\sigma_x)$ = standard deviation of $\sum X$

Calculator

To find probabilities for sums on the calculator, follow these steps.

2nd DISTR

2: normalcdf

normalcdf (lower value of the area, upper value of the area, $(n)(\text{mean})$, $(\sqrt{n})(\text{standard deviation})$)

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example 8.2.1

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

- Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7,500.
- Find the sum that is 1.5 standard deviations above the mean of the sums.

Answer

Let X = one value from the original unknown population. The probability question asks you to find a probability for the sum (or total of) 80 values.

$\sum X$ = the sum or total of 80 values. Since $\mu_x = 90$, $\sigma_x = 15$, and $n = 80$, $\sum X \sim N((80)(90), (\sqrt{80})(15))$

- mean of the sums = $(n)(\mu_x) = (80)(90) = 7,200$
- standard deviation of the sums = $(\sqrt{n})(\sigma_x) = (\sqrt{80})(15) = (80)(15)$
- sum of 80 values = $\sum X = 7,500$

a. Find $P(\sum X > 7,500)$

$$P(\sum X > 7,500) = 0.0127$$

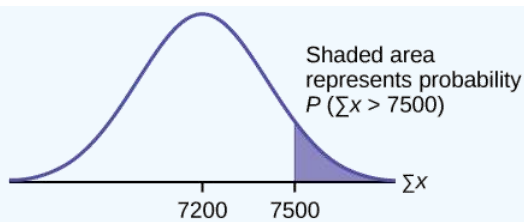


Figure 8.2.1

`normalcdf` (lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated (*lower, upper, (n)(μ_x, (√n)(σ_x))*)

`normalcdf` (7500, 1E99, (80)(90), (√80)(15)) = 0.0127

REMINDER

1E99 = 10⁹⁹.

Press the `EE` key for E.

b. Find $\sum x$ where $z = 1.5$.

$$\sum x = (n)(\mu_x) + (z)(\sqrt{n})(\sigma_x) = (80)(90) + (1.5)(\sqrt{80})(15) = 7,401.2$$

? Exercise 8.2.1

An unknown distribution has a mean of 45 and a standard deviation of eight. A sample size of 50 is drawn randomly from the population. Find the probability that the sum of the 50 values is more than 2,400.

Answer

0.0040

📊 Calculator

To find percentiles for sums on the calculator, follow these steps.

2nd DIST

3:invNorm

$k = \text{invNorm}(\text{area to the left of } k, (n)(\text{mean}), (\sqrt{n})(\text{standard deviation}))$

where:

- k is the k^{th} **percentile**
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

✓ Example 8.2.2

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. The sample of size is 50.

- What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
- Find the probability that the sum of the ages is between 1,500 and 1,800 years.
- Find the 80th percentile for the sum of the 50 ages.

Answer

a. $\mu_x - n\mu_x = 1,700$ and $\sigma_{\sum X} = \sqrt{n}\sigma_x = (\sqrt{50})(15) = 106.01$

The distribution is normal for sums by the central limit theorem.

b. $P(1500 < \sum X < 1800) = (1,500, 1,800, (50)(34), (\sqrt{50})(15)) = 0.7974$

- c. Let k = the 80th percentile.
 $k = (0.80, (50)(34), (\sqrt{50})(15)) = 1,789.3$

? Exercise 8.2.2

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 35 years. Suppose the standard deviation is ten years. The sample size is 39.

- What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
- Find the probability that the sum of the ages is between 1,400 and 1,500 years.
- Find the 90th percentile for the sum of the 39 ages.

Answer

- $\mu_{\sum X} = n\mu_X = 1,365$ and $\sigma_{\sum X} = \sqrt{n}\sigma_x = 62.4$
 The distribution is normal for sums by the central limit theorem.
- $P(1400 < \sum X < 1500) = \text{normalcdf}(1400, 1500, (39)(35), (\sqrt{39})(10)) = 0.2723$
- Let k = the 90th percentile.
 $k = \text{invNorm}(0.90, (39)(35), (\sqrt{39})(10)) = 1445.0$

✓ Example 8.2.3

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 size 70.

- What are the mean and standard deviation for the sums?
- Find the 95th percentile for the sum of the sample. Interpret this value in a complete sentence.
- Find the probability that the sum of the sample is at least ten hours.

Answer

- $\mu_{\sum X} = n\mu_X = 70(8.2) = 574$ minutes and $\sigma_{\sum X}(\sqrt{n})(\sigma_x) = (\sqrt{70})(1) = 8.37$ minutes
- Let k = the 95th percentile.
 $k = \text{invNorm}(0.95, (70)(8.2), (\sqrt{70})(1)) = 587.76$ minutes
 Ninety five percent of the app engagement times are at most 587.76 minutes.
- ten hours = 600 minutes
 $P(\sum X \geq 600) = \text{normalcdf}(600, E99, (70)(8.2), (\sqrt{70})(1)) = 0.0009$

? Exercise 8.2.3

The mean number of minutes for app engagement by a table use is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample size of 70.

- What is the probability that the sum of the sample is between seven hours and ten hours? What does this mean in context of the problem?
- Find the 84th and 16th percentiles for the sum of the sample. Interpret these values in context.

Answer

- 7 hours = 420 minutes
 10 hours = 600 minutes
 $\text{normalcdf}P(420 \leq \sum X \leq 600) = \text{normalcdf}(420, 600, (70)(8.2), \sqrt{70}(1)) = 0.9991$
 This means that for this sample sums there is a 99.9% chance that the sums of usage minutes will be between 420 minutes and 600 minutes.
- $\text{invNorm}(0.84, (70)(8.2), \sqrt{70}(1)) = 582.32$
 $\text{invNorm}(0.16, (70)(8.2), (\sqrt{70})(1)) = 565.68$
 Since 84% of the app engagement times are at most 582.32 minutes and 16% of the app engagement times are at most 565.68 minutes, we may state that 68% of the app engagement times are between 565.68 minutes and 582.32 minutes.

References

1. Farago, Peter. "The Truth About Cats and Dogs: Smartphone vs Tablet Usage Differences." The Flurry Blog, 2013. Posted October 29, 2012. Available online at blog.flurry.com (accessed May 17, 2013).

Review

The central limit theorem tells us that for a population with any distribution, the distribution of the sums for the sample means approaches a normal distribution as the sample size increases. In other words, if the sample size is large enough, the distribution of the sums can be approximated by a normal distribution even if the original population is not normally distributed. Additionally, if the original population has a mean of μ_x and a standard deviation of σ_x , the mean of the sums is $n\mu_x$ and the standard deviation is $(\sqrt{n})(\sigma_x)$ where n is the sample size.

Formula Review

- The Central Limit Theorem for Sums: $\sum X \sim N[(n)(\mu_x), (\sqrt{n})(\sigma_x)]$
- Mean for Sums $(\sum X) : (n)(\mu_x)$
- The Central Limit Theorem for Sums z -score and standard deviation for sums: z for the sample mean $= \frac{\sum x - (n)(\mu_x)}{(\sqrt{n})(\sigma_x)}$
- Standard deviation for Sums $(\sum X) : (\sqrt{n})(\sigma_x)$

This page titled [8.2: The Central Limit Theorem for Sums](#) 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.

8.3: Using the Central Limit Theorem

It is important for you to understand when to use the central limit theorem (clt). If you are being asked to find the probability of the mean, use the clt for the mean. If you are being asked to find the probability of a sum or total, use the clt for sums. This also applies to percentiles for means and sums.

If you are being asked to find the probability of an individual value, do not use the clt. Use the distribution of its random variable.

Law of Large Numbers

The law of large numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample tends to get closer and closer to μ . From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\bar{x}$ must be close to the population mean μ . We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

✓ Example 8.3.1

A study involving stress is conducted among the students on a college campus. The stress scores follow a uniform distribution with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students, find:

- The probability that the **mean stress score** for the 75 students is less than two.
- The 90th percentile for the **mean stress score** for the 75 students.
- The probability that the **total of the 75 stress scores** is less than 200.
- The 90th percentile for the **total stress score** for the 75 students.

Solutions

Let X = one stress score.

Problems a and b ask you to find a probability or a percentile for a mean. Problems c and d ask you to find a probability or a percentile for a **total or sum**. The sample size, n , is equal to 75.

Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$.

$$\mu_x = \frac{a+b}{2} = \frac{1+5}{2} = 3 \quad (8.3.1)$$

$$\sigma_x = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15 \quad (8.3.2)$$

For problems 1. and 2., let $\bar{X}$ = the mean stress score for the 75 students. Then,

$$\bar{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right) \quad (8.3.3)$$

where $n = 75$.

- Find $P(\bar{x} < 2)$. Draw the graph.
- Find the 90th percentile for the mean of 75 stress scores. Draw a graph.
- Find $P(\sum x < 2000)$. Draw the graph.
- Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Answers

a. $P(\bar{x} < 2) = 0$

The probability that the mean stress score is less than two is about zero.

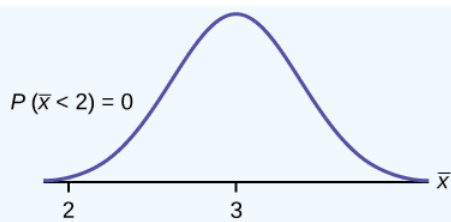


Figure 8.3.1

$$\text{normalcdf} \left(1, 2, 3, \frac{1.15}{\sqrt{75}} \right) = 0$$

REMINDER

The smallest stress score is one

b. Let k = the 90th percentile.

Find k , where $P(\bar{x} < k) = 0.90$.

$$k = 3.2$$

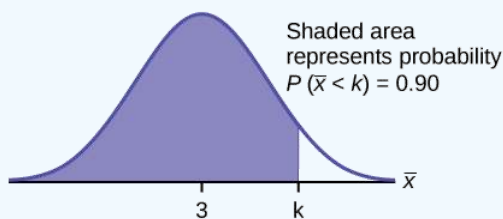


Figure 8.3.2

The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2, and that 10% are at least 3.2.

$$\text{invNorm} \left(0.90, 3, 1, \frac{1.15}{\sqrt{75}} \right) = 3.2$$

For problems c and d, let $\sum X$ = the sum of the 75 stress scores. Then,

$$\sum X \sim N((75)(3), (\sqrt{75})(1.15)) \quad (8.3.4)$$

c. The mean of the sum of 75 stress scores is $(75)(3) = 225$

The standard deviation of the sum of 75 stress scores is $(\sqrt{75})(1.15) = 9.96$

$$P(\sum x < 200)$$

$$P(\sum x < 200) = 0$$

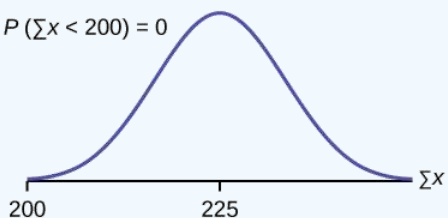


Figure 8.3.3

The probability that the total of 75 scores is less than 200 is about zero.

$$\text{normalcdf} \quad 75, 200, (75)(3), (\sqrt{75})(1.15)$$

REMINDER

The smallest total of 75 stress scores is 75, because the smallest single score is one.

d. Let k = the 90th percentile.

Find k where $P(\sum x < k) = 0.90$.

$$k = 237.8$$

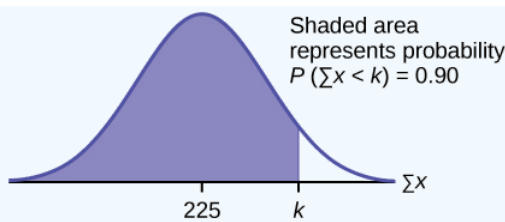


Figure 8.3.4

The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

$$\text{invNorm}(0.90, (75)(3), (\sqrt{75})(1.15)) = 237.8$$

? Exercise 8.3.1

Use the information in Example 8.3.1, but use a sample size of 55 to answer the following questions.

- Find $P(\bar{x} < 7)$.
- Find $P(\sum x < 7)$.
- Find the 80th percentile for the mean of 55 scores.
- Find the 85th percentile for the sum of 55 scores.

Answer

- 0.0265
- 0.2789
- 3.13
- 173.84

✓ Example 8.3.2

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the *excess time used* follows an exponential distribution with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim \text{Exp}\left(\frac{1}{22}\right)$. From previous chapters, we know that $\mu = 22$ and $\sigma = 22$.

Let $\bar{X}$ = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$$\bar{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right) \quad (8.3.5)$$

by the central limit theorem for sample means

- Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(\bar{x} > 20)$. Draw the graph.
- Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$.
- Explain why the probabilities in parts a and b are different.
- Find the 95th percentile for the **sample mean excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Answer

- Find: $P(\bar{x} > 20)$

$$P(\bar{x} > 20) = 0.79199 \text{ using } \text{normalcdf} \left(20, 1E99, 22, \frac{22}{\sqrt{80}} \right)$$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.

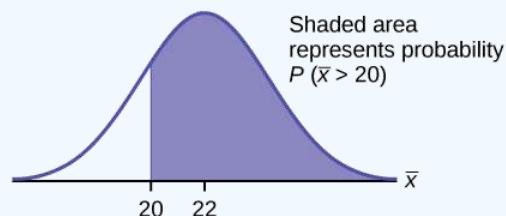


Figure 8.3.5.

REMINDER

1E99 = 10^{99} and $-1E99 = -10^{99}$. Press the **EE** key for E. Or just use 10^{99} instead of 1E99.

- b. Find $P(x > 20)$. Remember to use the exponential distribution for an **individual**: $X \sim \text{Exp} \left(\frac{1}{22} \right)$.

$$P(x > 20) = e^{-\left(\frac{1}{22}\right)(20)} \text{ or } e^{(-0.04545)(20)} = 0.4029$$

- c. i. $P(x > 20) = 0.4029$ but $P(\bar{x} > 20) = 0.7919$
 ii. The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.
 iii. When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the clt. Use the clt with the normal distribution when you are being asked to find the probability for a mean.
 d. Let k = the 95th percentile. Find k where $P(\bar{x} < k) = 0.95$

$$k = 26.0 \text{ using } \text{invNorm} \left(0.95, 22, \frac{22}{\sqrt{80}} \right) = 26.0$$

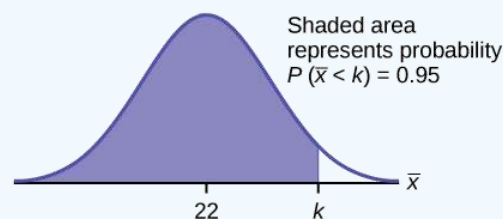


Figure 8.3.6.

The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

Ninety five percent of such samples would have means under 26 minutes; only five percent of such samples would have means above 26 minutes.

? Exercise 8.3.2

Use the information in Example 8.3.2, but change the sample size to 144.

- Find $P(20 < \bar{x} < 30)$.
- Find $P(\sum x \text{ is at least } 3,000)$.
- Find the 75th percentile for the sample mean excess time of 144 customers.
- Find the 85th percentile for the sum of 144 excess times used by customers.

Answer

- a. 0.8623
- b. 0.7377
- c. 23.2
- d. 3,441.6

✓ Example 8.3.3

In the United States, someone is sexually assaulted every two minutes, on average, according to a number of studies. Suppose the standard deviation is 0.5 minutes and the sample size is 100.

- a. Find the median, the first quartile, and the third quartile for the sample mean time of sexual assaults in the United States.
- b. Find the median, the first quartile, and the third quartile for the sum of sample times of sexual assaults in the United States.
- c. Find the probability that a sexual assault occurs on the average between 1.75 and 1.85 minutes.
- d. Find the value that is two standard deviations above the sample mean.
- e. Find the *IQR* for the sum of the sample times.

Answer

- a. We have, $\mu_x = \mu = 2$ and $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$. Therefore:
 - a. 50th percentile = $\mu_x = \mu = 2$
 - b. 25th percentile = $\text{invNorm}(0.25, 2, 0.05) = 1.97$
 - c. 75th percentile = $\text{invNorm}(0.75, 2, 0.05) = 2.03$
- b. We have $\mu_{\sum X} = n(\mu_x) = 100(2)$ and $\sigma_{\sum X} = \sqrt{n}(\sigma_x) = 10(0.5) = 5$. Therefore
 - a. 50th percentile = $\mu_{\sum X} = n(\mu_x) = 100(2) = 200$
 - b. 25th percentile = $\text{invNorm}(0.25, 200, 5) = 196.63$
 - c. 75th percentile = $\text{invNorm}(0.75, 200, 5) = 203.37$
- c. $P(1.75 < \bar{x} < 1.85) = \text{normalcdf}(1.75, 1.85, 2, 0.05) = 0.0013$
- d. Using the *z*-score equation, $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$, and solving for *x*, we have $x = 2(0.05) + 2 = 2.1$
- e. The *IQR* is 75th percentile – 25th percentile = $203.37 - 196.63 = 6.74$

? Exercise 8.3.3

Based on data from the National Health Survey, women between the ages of 18 and 24 have an average systolic blood pressures (in mm Hg) of 114.8 with a standard deviation of 13.1. Systolic blood pressure for women between the ages of 18 to 24 follow a normal distribution.

- a. If one woman from this population is randomly selected, find the probability that her systolic blood pressure is greater than 120.
- b. If 40 women from this population are randomly selected, find the probability that their mean systolic blood pressure is greater than 120.
- c. If the sample were four women between the ages of 18 to 24 and we did not know the original distribution, could the central limit theorem be used?

Answer

- a. $P(x > 120) = \text{normalcdf}(120, 99, 114.8, 13.1) = 0.0272$ There is about a 3%, that the randomly selected woman will have systolics blood pressure greater than 120.
- b. $P(\bar{x} > 120) = \text{normalcdf}\left(120, 114.8, \frac{13.1}{\sqrt{40}}\right) = 0.006$ There is only a 0.6% chance that the average systolic blood pressure for the randomly selected group is greater than 120.
- c. The central limit theorem could not be used if the sample size were four and we did not know the original distribution was normal. The sample size would be too small.

✓ Example 8.3.4

A study was done about violence against prostitutes and the symptoms of the posttraumatic stress that they developed. The age range of the prostitutes was 14 to 61. The mean age was 30.9 years with a standard deviation of nine years.

- In a sample of 25 prostitutes, what is the probability that the mean age of the prostitutes is less than 35?
- Is it likely that the mean age of the sample group could be more than 50 years? Interpret the results.
- In a sample of 49 prostitutes, what is the probability that the sum of the ages is no less than 1,600?
- Is it likely that the sum of the ages of the 49 prostitutes is at most 1,595? Interpret the results.
- Find the 95th percentile for the sample mean age of 65 prostitutes. Interpret the results.
- Find the 90th percentile for the sum of the ages of 65 prostitutes. Interpret the results.

Answer

- $P(\bar{x} < 35) = \text{normalcdf}(-E99, 35, 30.9, 1.8) = 0.9886$
- $P(\bar{x} > 50) = \text{normalcdf}(50, E99, 30.9, 1.8) \approx 0$ For this sample group, it is almost impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.
- $P(\sum x \geq 1,600) = \text{normalcdf}(1600, E99, 1514.10, 63) = 0.0864$
- $P(\sum x \leq 1,595) = \text{normalcdf}(-E99, 1595, 1514.10, 63) = 0.9005$ This means that there is a 90% chance that the sum of the ages for the sample group $n = 49$ is at most 1595.
- The 95th percentile = $\text{invNorm}(0.95, 30.9, 1.1) = 32.7$ This indicates that 95% of the prostitutes in the sample of 65 are younger than 32.7 years, on average.
- The 90th percentile = $\text{invNorm}(0.90, 2008.5, 72.56) = 2101.5$ This indicates that 90% of the prostitutes in the sample of 65 have a sum of ages less than 2,101.5 years.

? Exercise 8.3.4

According to Boeing data, the 757 airliner carries 200 passengers and has doors with a mean height of 72 inches. Assume for a certain population of men we have a mean of 69.0 inches and a standard deviation of 2.8 inches.

- What mean doorway height would allow 95% of men to enter the aircraft without bending?
- Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?
- For engineers designing the 757, which result is more relevant: the height from part a or part b? Why?

Answer

- We know that $\mu_x = \mu = 69$ and we have $\sigma_x = 2.8$. The height of the doorway is found to be $\text{invNorm}(0.95, 69, 2.8) = 73.61$
- We know that $\mu_x = \mu = 69$ and we have $\sigma_x = 2.8$. So, $\text{invNorm}(0.95, 69, 0.28) = 69.49$
- When designing the doorway heights, we need to incorporate as much variability as possible in order to accommodate as many passengers as possible. Therefore, we need to use the result based on part a.

📌 Historical Note: Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the central limit theorem. Binomial probabilities with a small value for n (say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n , you had to use the binomial formula, which could be very complicated. Using the normal approximation to the binomial distribution simplified the process. To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the conditions for a binomial distribution:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$); the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by

the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the continuity correction factor and is used in the following example.

✓ Example 8.3.5

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

- Find the probability that **at least 150** favor a charter school.
- Find the probability that **at most 160** favor a charter school.
- Find the probability that **more than 155** favor a charter school.
- Find the probability that **fewer than 147** favor a charter school.
- Find the probability that **exactly 175** favor a charter school.

Let X = the number that favor a charter school for grades K through 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is X . $Y \sim N(159, 8.6447)$ See The Normal Distribution for help with calculator instructions.

For part a, you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.

`normalcdf (149.5, 1099, 159, 8.6447) = 0.8641`

For part b, you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.

`normalcdf (0, 160.5, 159, 8.6447) = 0.5689`

For part c, you **exclude 155** so $P(X > 155)$ has normal approximation $P(y > 155.5) = 0.6572$

`normalcdf (155.5, 1099, 159, 8.6447) = 0.6572`

For part d, you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.

`normalcdf (0, 146.5, 159, 8.6447) = 0.0741`

For part e, $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$

`normalcdf (174.5, 175.5, 159, 8.6447) = 0.0083`

Because of calculators and computer software that let you calculate binomial probabilities for large values of n easily, it is not necessary to use the the normal approximation to the binomial distribution, provided that you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators, and they easily calculate probabilities for the binomial distribution. If you type in "binomial probability distribution calculation" in an Internet browser, you can find at least one online calculator for the binomial.

For Example, the probabilities are calculated using the following binomial distribution: ($n = 300$ and $p = 0.53$). Compare the binomial and normal distribution answers. See Discrete Random Variables for help with calculator instructions for the binomial.

$P(X \geq 150) : 1 - \text{binomialcdf} (300, 0.53, 149) = 0.8641$

$P(X \leq 160) : \text{binomialcdf} (300, 0.53, 160) = 0.5684$

$P(X > 155) : 1 - \text{binomialcdf} (300, 0.53, 155) = 0.6576$

$P(X < 147) : \text{binomialcdf} (300, 0.53, 146) = 0.0742$

$P(X = 175) : (\text{You use the binomial pdf.}) \text{binomialpdf} (300, 0.53, 175) = 0.0083$

? Exercise 8.3.5

In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

Answer

0.0401

References

- Data from the Wall Street Journal.
- "National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at <http://www.cdc.gov/nchs/nhanes.htm> (accessed May 17, 2013).

Glossary

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital, notation: $X \sim \text{Exp}(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Mean

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

Normal Distribution

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 the **standard normal distribution**.

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, $(a < x < b)$; often referred as the **Rectangular Distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

This page titled 8.3: Using the Central Limit Theorem 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.

- 7.4: Using the Central Limit Theorem by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/introductory-statistics>.

CHAPTER OVERVIEW

9: Confidence Intervals

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

[9.0: Prelude to Confidence Intervals](#)

[9.1: A Single Population Mean using the Normal Distribution](#)

[9.2: A Single Population Mean using the Student t-Distribution](#)

[9.3: A Population Proportion](#)

[9.S: Confidence Intervals \(Summary\)](#)

This page titled [9: Confidence Intervals](#) 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.

9.0: Prelude to Confidence Intervals

Learning Objectives

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

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's t distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.

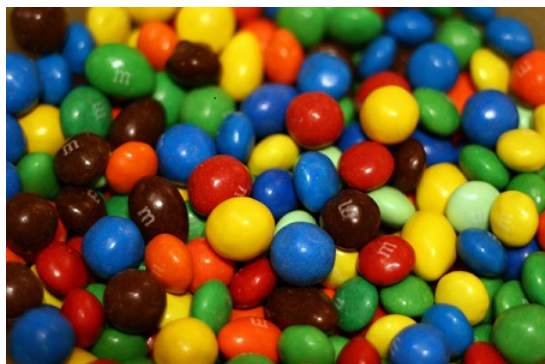


Figure 9.0.1 Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's- t , and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, s . You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ . The sample standard deviation, s , is the point estimate for the population standard deviation, σ .

Each of $\bar{x}$ and s is called a statistic.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ , but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1. \quad (9.0.1)$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ . For our iTunes example, two standard deviations is $(2)(0.1) = 0.2$. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ .

Because $\bar{x}$ is within 0.2 units of μ , which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations $(2)(0.1)$ and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \quad (9.0.2)$$

and

$$\bar{x} + 0.2 = 2 + 0.2 = 2.2 \quad (9.0.3)$$

We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is (1.8, 2.2). This 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ . The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, $\bar{x}$. Confidence intervals for some parameters have the form:

$$(\text{point estimate} - \text{margin of error}, \text{point estimate} + \text{margin of error})$$

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Collaborative Exercise

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval $\left(\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}} \right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between _____ and _____.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of

the production is defective.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

This page titled [9.0: Prelude to Confidence Intervals](#) 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.

- [8.1: Prelude to Confidence Intervals](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

9.1: A Single Population Mean using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ , where the population standard deviation is known, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated EBM). The sample mean $\bar{x}$ is the point estimate of the unknown population mean μ .

The confidence interval estimate will have the form:

$$(\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})$$

or, in symbols,

$$(\bar{x} - EBM, \bar{x} + EBM)$$

The **margin of error** (EBM) depends on the confidence level (abbreviated CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL . α is the probability that the interval does not contain the unknown population parameter. Mathematically,

$$\alpha + CL = 1.$$

✓ Example 9.1.1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5: $\bar{x} = 7$ and $EBM = 2.5$

The confidence interval is $(7 - 2.5, 7 + 2.5)$ and calculating the values gives $(4.5, 9.5)$. If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

? Exercise 9.1.1

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2. What is the confidence interval estimate for the population mean?

Answer

$(11.8, 18.2)$

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$. To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10$ in both tails, or 5% in each tail, of the normal distribution.

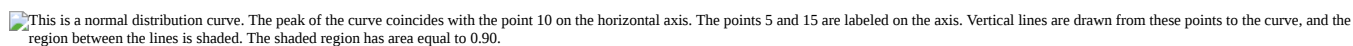
 This is a normal distribution curve. The peak of the curve coincides with the point 10 on the horizontal axis. The points 5 and 15 are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 0.90.

Figure 9.1.1

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is

$$\frac{\sigma}{\sqrt{n}}$$

This fraction is commonly called the "standard error of the mean" to distinguish clearly the standard deviation for a mean from the population standard deviation σ .

In summary, as a result of the central limit theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N(\mu_x, \frac{\sigma}{\sqrt{n}})$.
- When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ .
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM .
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ , we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL , is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

using a calculator, computer or a standard normal probability table.

`invNorm (0.975, 0, 1) = 1.96`

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0, 1)$.

Calculating the Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$EBM = z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Constructing the Confidence Interval

The confidence interval estimate has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

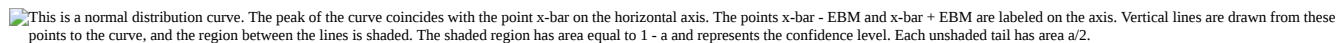
 This is a normal distribution curve. The peak of the curve coincides with the point $\bar{x}$ -bar on the horizontal axis. The points $\bar{x}$ -bar - EBM and $\bar{x}$ -bar + EBM are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to $1 - \alpha$ and represents the confidence level. Each unshaded tail has area $\alpha/2$.

Figure 8.2.2.

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ____% confidence that the true population mean (include the context of the problem) is between ____ and ____ (include appropriate units)."

✓ Example 9.1.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM .

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$\sigma = 3; n = 36$$

The confidence level is 90% ($CL = 0.90$)

$$CL = 0.90$$

so

$$\alpha = 1 - CL = 1 - 0.90 = 0.10$$

$$\frac{\alpha}{2} = 0.05 \quad z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1 - 0.05 = 0.95$.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

using $\text{invNorm}(0.95, 0, 1)$ on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$$EBM = (1.645) \left(\frac{3}{\sqrt{36}} \right) = 0.8225$$

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is **(67.1775, 68.8225)**.

Solution B

Press **STAT** and arrow over to **TESTS** .

Arrow down to **7:ZInterval** .

Press **ENTER** .

Arrow to **Stats** and press **ENTER** .

Arrow down and enter three for σ , 68 for $\bar{x}$, 36 for n , and .90 for **C-level** .

Arrow down to **Calculate** and press **ENTER** .

The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

? Exercise 9.1.2

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.1347, 37.8653)

✓ Example 9.1.3: Specific Absorption Rate

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messenger III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean.

$$\bar{x} = 1.024$$

Next, find the *EBM*. Because you are creating a 98% confidence interval, $CL = 0.98$.

 This is a normal distribution curve. The point $z_{0.01}$ is labeled at the right edge of the curve and the region to the right of this point is shaded. The area of this shaded region equals 0.01. The unshaded area equals 0.99.

Figure 8.2.3.

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01}) \frac{\sigma}{\sqrt{n}} = (2.326) \frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$$\bar{x} - EBM = 1.024 - 0.1431 = 0.8809$$

$$\bar{x} + EBM = 1.024 + 0.1431 = 1.1671$$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to 7:Z Interval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:
 - σ : 0.337
 - $\bar{x}$: 1024
 - n : 30
 - C-level: 0.98
- Arrow down to Calculate and press ENTER.
- The confidence interval is (to three decimal places) (0.881, 1.167).

? Exercise 9.1.3

Table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Answer

$$\bar{x} = 0.940$$

$$\frac{\alpha}{2} = \frac{1 - CL}{2} = \frac{1 - 0.93}{2} = 0.035$$

$$z_{0.035} = 1.812$$

$$EBM = (z_{0.035}) \left(\frac{\sigma}{\sqrt{n}} \right) = (1.812) \left(\frac{0.337}{\sqrt{20}} \right) = 0.1365$$

$$\bar{x} - EBM = 0.940 - 0.1365 = 0.8035$$

$$\bar{x} + EBM = 0.940 + 0.1365 = 1.0765$$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in Example and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

✓ Example 9.1.4

Suppose we change the original problem in Example by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM .

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right)$$

$\sigma = 3; n = 36$; The confidence level is 95% ($CL = 0.95$).

$$CL = 0.95 \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 \quad z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

when using $\text{invnorm}(0.975, 0, 1)$ on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96) \left(\frac{3}{\sqrt{36}} \right) = 0.98$$

$$\bar{x} - EBM = 68 - 0.98 = 67.02$$

$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the EBM is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

 Part (a) shows a normal distribution curve. A central region with area equal to 0.90 is shaded. Each unshaded tail of the curve has area equal to 0.05. Part (b) shows a normal distribution curve. A central region with area equal to 0.95 is shaded. Each unshaded tail of the curve has area equal to 0.025.

Figure 8.2.4.

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

? Exercise 9.1.4

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Answer

(33.37, 38.63)

✓ Example 9.1.5

Suppose we change the original problem in Example to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right)$
- $\sigma = 3$; The confidence level is 90% ($CL=0.90$); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Answer

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.

$$\text{When } n = 100 : EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right) = (1.645) \left(\frac{3}{\sqrt{100}} \right) = 0.4935.$$

Solution B

If we **decrease** the sample size n to 25, we **increase** the error bound.

$$\text{When } n = 25 : EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right) = (1.645) \left(\frac{3}{\sqrt{25}} \right) = 0.987.$$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

? Exercise 9.1.5

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

✓ Example 9.1.6

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68 : $EBM = 68.82 - 68 = 0.82$.
- If we don't know the sample mean: $EBM = \frac{(68.82 - 67.18)}{2} = 0.82$.

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

? Exercise 9.1.6

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Answer

Sample mean is 45, error bound is 2.88

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population mean when the population standard deviation is known is

$$EBM = \left(z \frac{\sigma}{\sqrt{n}} \right)$$

The formula for sample size is $n = \frac{z^2 \sigma^2}{EBM^2}$, found by solving the error bound formula for n . In Equation ???, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

✓ Example 9.1.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

Solution

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2 \sigma^2}{EBM^2} = \frac{(1.96)^2 (15)^2}{2^2}$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

? Exercise 9.1.7

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Answer

35 students

References

1. "American Fact Finder." U.S. Census Bureau. Available online at <http://factfinder2.census.gov/faces/...html?refresh=t> (accessed July 2, 2013).
2. "Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2, 2013).
3. "Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at research.fhda.edu/factbook/FH...phicTrends.htm (accessed September 30, 2013).
4. Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at www.cdc.gov/growthcharts/2000...thchart-us.pdf (accessed July 2, 2013).
5. La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at <http://reviews.cnet.com/cell-phone-radiation-levels/> (accessed July 2, 2013).
6. "Mean Income in the Past 12 Months (in 2011 Inflation-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at <http://factfinder2.census.gov/faces/...prodType=table> (accessed July 2, 2013).
7. "Metadata Description of Candidate Summary File." U.S. Federal Election Commission. Available online at www.fec.gov/finance/disclosure...esummary.shtml (accessed July 2, 2013).
8. "National Health and Nutrition Examination Survey." Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/nchs/nhanes.htm> (accessed July 2, 2013).

Glossary

Confidence Level (CL)

the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the $CL = 90$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Mean (EBM)

the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

This page titled [9.1: A Single Population Mean using the Normal Distribution](#) 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.

- [8.2: A Single Population Mean using the Normal Distribution](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

9.2: A Single Population Mean using the Student t-Distribution

In practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's t -distribution. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's t -distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t -distribution whenever s is used as an estimate for σ . If you draw a simple random sample of size n from a population that has an approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t -score

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}, \quad (9.2.1)$$

then the t -scores follow a Student's t -distribution with $n-1$ degrees of freedom. The t -score has the same interpretation as the z -score. It measures how far $\bar{x}$ is from its mean μ . For each sample size n , there is a different Student's t -distribution.

The degrees of freedom, $n-1$, come from the calculation of the sample standard deviation s . Previously, we used n deviations ($x - \bar{x}$ values) to calculate s . Because the sum of the deviations is zero, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. We call the number $n-1$ the degrees of freedom (df).

For each sample size n , there is a different Student's t -distribution.

Properties of the Student's t -Distribution

- The graph for the Student's t -distribution is similar to the standard normal curve.
- The mean for the Student's t -distribution is zero and the distribution is symmetric about zero.
- The Student's t -distribution has more probability in its tails than the standard normal distribution because the spread of the t -distribution is greater than the spread of the standard normal. So the graph of the Student's t -distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t -distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t -distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ . The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t -probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of t . The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t -score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's t -distribution can also be used. The table gives t -scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator,

you need to use a probability table for the Student's t -Distribution.) When using a t -table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's t -table gives t -scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t -probabilities.**

The notation for the Student's t -distribution (using T as the random variable) is:

- $T \sim t_{df}$ where $df = n - 1$.
- For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

If the population standard deviation is not known, the error bound for a population mean is:

- $EBM = \left(t_{\frac{\alpha}{2}} \right) \left(\frac{s}{\sqrt{n}} \right)$,
- $t_{\frac{\alpha}{2}}$ is the t -score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is:

$$(\bar{x} - EBM, \bar{x} + EBM). \quad (9.2.2)$$

To calculate the confidence interval directly:

Press STAT.

Arrow over to TESTS.

Arrow down to 8:TInterval and press ENTER (or just press 8).

✓ Example 9.2.1: Acupuncture

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM .

$$\bar{x} = 8.2267$$

$$s = 1.6722$$

$$n = 15$$

$$df = 15 - 1 = 14$$

$$CL \text{ is } 95\%, \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$t_{\frac{\alpha}{2}} = t_{0.025}$$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is $1 - 0.025 = 0.975$

$$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14 \text{ using invT(.975,14) on the TI-84+ calculator.}$$

$$EBM = \left(t_{\frac{\alpha}{2}}\right) \left(\frac{s}{\sqrt{n}}\right)$$

$$= (2.14) \left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

Now it is just a direct application of Equation 9.2.2:

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$

$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Press **STAT** and arrow over to **TESTS**.

Arrow down to **8: TInterval** and press **ENTER** (or you can just press **8**).

Arrow to **Data** and press **ENTER**.

Arrow down to **List** and enter the list name where you put the data.

There should be a 1 after **Freq**.

Arrow down to **C-level** and enter 0.95

Arrow down to **Calculate** and press **ENTER**.

The 95% confidence interval is (7.3006, 9.1527)

When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t . The table gives t -scores that correspond to the confidence level (column) and degrees of freedom (row); the t -score is found where the row and column intersect in the table.

? Exercise 9.2.1

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Answer

(8.1634, 9.8032)

✓ Example 9.2.2: The Human Toxome Project

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. Table 9.2.1 shows how many of the targeted chemicals were found in each infant's cord blood.

Table 9.2.1

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an infant's blood.

Solution A

From the sample, you can calculate $\bar{x} = 127.45$ and $s = 25.965$. There are 20 infants in the sample, so $n = 20$, and $df = 20 - 1 = 19$.

You are asked to calculate a 90% confidence interval: $CL = 0.90$, so

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \quad (9.2.3)$$

$$\frac{\alpha}{2} = 0.05$$

$$t_{\frac{\alpha}{2}} = t_{0.05} \quad (9.2.4)$$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is $1 - 0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 1.729 \left(\frac{25.965}{\sqrt{20}} \right) \approx 10.038$$

$$\bar{x} - EBM = 127.45 - 10.038 = 117.412$$

$$\bar{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution B

Enter the data as a list.

Press **STAT** and arrow over to **TESTS**.

Arrow down to **8:TInterval** and press **ENTER** (or you can just press **8**). Arrow to **Data** and press **ENTER**.

Arrow down to **List** and enter the list name where you put the data.

Arrow down to **Freq** and enter 1.

Arrow down to **C-level** and enter 0.90.

Arrow down to **Calculate** and press **ENTER**.

The 90% confidence interval is (117.41, 137.49).

✓ Example 9.2.3

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in Table 9.2.2. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

Table 9.2.2

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution A

- $\bar{x} = 6.133$,
- $s = 5.514$,
- $n = 15$, and
- $df = 15 - 1 = 14$.

$$CL = 0.98, \text{ so } \alpha = 1 - CL = 1 - 0.98 = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

$$t_{\frac{\alpha}{2}} = t_{0.01} = 2.624$$

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 2.624 \left(\frac{5.514}{\sqrt{15}} \right) = 3.736$$

$$\bar{x} - EBM = 6.133 - 3.736 = 2.397$$

$$\bar{x} + EBM = 6.133 + 3.736 = 9.869$$

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution B

Enter the data as a list.

Press **STAT** and arrow over to **TESTS**.

Arrow down to **8:TInterval**.

Press **ENTER**.

Arrow to **Data** and press **ENTER**.

Arrow down and enter the name of the list where the data is stored.

Enter **Freq : 1**

Enter **C-Level : 0.98**

Arrow down to **Calculate** and press **Enter**.

The 98% confidence interval is (2.3965, 9.8702).

Reference

1. "America's Best Small Companies." Forbes, 2013. Available online at <http://www.forbes.com/best-small-companies/list/> (accessed July 2, 2013).
2. Data from *Microsoft Bookshelf*.
3. Data from <http://www.businessweek.com/>.
4. Data from <http://www.forbes.com/>.
5. "Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012." Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2, 2013).
6. "Human Toxome Project: Mapping the Pollution in People." Environmental Working Group. Available online at www.ewg.org/sites/humantoxome...tero%2Fnewborn (accessed July 2, 2013).
7. "Metadata Description of Leadership PAC List." Federal Election Commission. Available online at www.fec.gov/finance/disclosur...pPacList.shtml (accessed July 2, 2013).

Glossary

Degrees of Freedom (df)

the number of objects in a sample that are free to vary

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, 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 **the standard normal distribution**.

Standard Deviation

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

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n get larger.
- There is a "family" of t-distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

This page titled [9.2: A Single Population Mean using the Student t-Distribution](#) 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.

- [8.3: A Single Population Mean using the Student t-Distribution](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

9.3: A Population Proportion

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: $(0.40 - 0.03, 0.40 + 0.03)$.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the error bound, and the confidence level for a proportion is similar to that for the population mean, but the formulas are different. How do you know you are dealing with a proportion problem? First, the underlying distribution is a binomial distribution. (There is no mention of a mean or average.) If X is a binomial random variable, then

$$X \sim B(n, p)$$

where n is the number of trials and p is the probability of a success.

To form a proportion, take X , the random variable for the number of successes and divide it by n , the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\hat{P}$, read "P hat".)

When n is large and p is not close to zero or one, we can use the normal distribution to approximate the binomial.

$$X \sim N(np, \sqrt{npq})$$

If we divide the random variable, the mean, and the standard deviation by n , we get a normal distribution of proportions with P' , called the estimated proportion, as the random variable. (Recall that a proportion is the number of successes divided by n .)

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Using algebra to simplify:

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

P' follows a normal distribution for proportions:

$$\frac{X}{n} = P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

The confidence interval has the form

$$(p' - EBP, p' + EBP).$$

where

- EBP is error bound for the proportion.
- $p' = \frac{x}{n}$
- p' = the estimated proportion of successes (p' is a point estimate for p , the true proportion.)
- x = the number of successes
- n = the size of the sample

The error bound (EBP) for a proportion is

$$EBP = \left(z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{p'q'}{n}} \right)$$

where $q' = 1 - p'$.

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is

$$\sqrt{\frac{pq}{n}}.$$

However, in the error bound formula, we use

$$\sqrt{\frac{p'q'}{n}}$$

as the standard deviation, instead of

$$\sqrt{\frac{pq}{n}}.$$

In the error bound formula, the sample proportions p' and q' are estimates of the unknown population proportions p and q . The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

Normal Distribution of Proportions

For the normal distribution of proportions, the z -score formula is as follows.

If

$$P' \sim N \left(p, \sqrt{\frac{pq}{n}} \right) \quad (9.3.1)$$

then the z -score formula is

$$z = \frac{p' - p}{\sqrt{\frac{pq}{n}}} \quad (9.3.2)$$

✓ Example 9.3.1

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution A

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Let X = the number of people in the sample who have cell phones. X is binomial.

$$X \sim B(500, \frac{421}{500}).$$

To calculate the confidence interval, you must find p' , q' , and EBP .

- $n = 500$
- $x = \text{the number of successes} = 421$

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

- $p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158$$

Since $CL = 0.95$, then

$$\alpha = 1 - CL = 1 - 0.95 = 0.05 \left(\frac{\alpha}{2} \right) = 0.025.$$

Then

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

Use the TI-83, 83+, or 84+ calculator command $\text{invNorm}(0.975, 0, 1)$ to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.96) \sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.810, 0.874)$

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution B

Press **STAT** and arrow over to **TESTS**.

Arrow down to **A:1-PropZint**. Press **ENTER**.

Arrow down to **xx** and enter 421.

Arrow down to **nn** and enter 500.

Arrow down to **C-Level** and enter .95.

Arrow down to **Calculate** and press **ENTER**.

The confidence interval is (0.81003, 0.87397).

? Exercise 9.3.1

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Answer

(0.3315, 0.4525)

✓ Example 9.3.2

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

- $x = 300$ and
- $n = 500$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \left(\frac{\alpha}{2} \right) = 0.05 \quad (9.3.3)$$

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command `invNorm(0.95,0,1)` to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}} \right) \sqrt{\frac{p'q'}{n}} = (1.645) \sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.564, 0.636)$.

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Press `STAT` and arrow over to `TESTS`.

Arrow down to `A:1-PropZint`. Press `ENTER`.

Arrow down to `xx` and enter 300.

Arrow down to `nn` and enter 500.

Arrow down to `C-Level` and enter 0.90.

Arrow down to `Calculate` and press `ENTER`.

The confidence interval is (0.564, 0.636).

? Exercise 9.3.2

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

- Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.
- In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Answer a

(0.7731, 0.8269); We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.31% and 82.69%.

Answer b

Sixty-eight percent (68%) of students own an iPod and a smart phone.

$$p' = 0.68$$

$$q' = 1 - p' = 1 - 0.68 = 0.32$$

Since $CL = 0.97$, we know

$$\alpha = 1 - 0.97 = 0.03$$

and

$$\frac{\alpha}{2} = 0.015.$$

The area to the left of $z_{0.05}$ is 0.015, and the area to the right of $z_{0.05}$ is $1 - 0.015 = 0.985$.

Using the TI 83, 83+, or 84+ calculator function $\text{InvNorm}(0.985, 0, 1)$,

$$z_{0.05} = 2.17$$

$$EPB = \left(z_{\frac{\alpha}{2}} \right) \sqrt{\frac{p'q'}{n}} = 2.17 \sqrt{\frac{0.68(0.32)}{300}} \approx 0.0269$$

$$p' - EPB = 0.68 - 0.0269 = 0.6531$$

$$p' + EPB = 0.68 + 0.0269 = 0.7069$$

We are 97% confident that the true proportion of all students who own an iPod and a smart phone is between 0.6531 and 0.7069.

Calculator

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Arrow down to x and enter $300 \cdot 0.68$.

Arrow down to n and enter 300.

Arrow down to C-Level and enter 0.97.

Arrow down to Calculate and press ENTER.

The confidence interval is (0.6531, 0.7069).

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population proportion is

$$EBP = \left(z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{p'q'}{n}} \right)$$

Solving for n gives you an equation for the sample size.

$$n = \frac{\left(z_{\frac{\alpha}{2}} \right)^2 (p'q')}{EBP^2}$$

✓ Example 9.3.3

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Answer

From the problem, we know that **EBP** = **0.03** (3%=0.03) and $z_{\frac{\alpha}{2}} \approx z_{0.05} = 1.645$ because the confidence level is 90%.

However, in order to find n , we need to know the estimated (sample) proportion p' . Remember that $q' = 1 - p'$. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q' = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n . This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n , use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{EBP^2}$$

gives

$$n = \frac{1.645^2 (0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

? Exercise 9.3.3

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Answer

271 customers should be surveyed. Check the Real Estate section in your local

Glossary

Binomial Distribution

a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n , of independent trials.

“Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all

trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Error Bound for a Population Proportion (*EBP*)

the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

This page titled [9.3: A Population Proportion](#) 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.

- **8.4: A Population Proportion** by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

9.S: Confidence Intervals (Summary)

Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

$$(\text{lower bound}, \text{upper bound}) = (\text{point estimate} - EBM, \text{point estimate} + EBM)$$

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = z \frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for *n* to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2 \sigma^2}{EBM^2}$$

Formula Review

$\bar{X} - N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

$$(\text{lower bound}, \text{upper bound}) = (\text{point estimate} - EBM, \text{point estimate} + EBM) \quad (9.S.1)$$

$$= \bar{x} - EBM, \bar{x} + EBM \quad (9.S.2)$$

$$= \left(\bar{x} - z \frac{\sigma}{\sqrt{n}}, \bar{x} + z \frac{\sigma}{\sqrt{n}} \right) \quad (9.S.3)$$

$EBM = z \frac{\sigma}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the *z*-score with the property that the area to the right of the *z*-score is $\frac{\alpha}{2}$ this is the *z*-score used in the calculation of "*EBM*" where $\alpha = 1 - CL$.

$n = \frac{z^2 \sigma^2}{EBM^2}$ the formula used to determine the sample size (*n*) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

$$(\text{lower value}, \text{upper value}) = (\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound}) \quad (9.S.4)$$

To find the error bound when you know the confidence interval

$$\text{error bound} = \text{upper value} - \text{point estimate} \quad (9.S.5)$$

OR

$$\text{error bound} = \frac{\text{upper value} - \text{lower value}}{2} \quad (9.S.6)$$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

? Exercise 8.2.8

Identify the following:

- a. $\bar{x}$ = _____
- b. σ = _____
- c. n = _____

Answer

- a. 244
- b. 15
- c. 50

? Exercise 8.2.9

In words, define the random variables X and $\bar{X}$.

? Exercise 8.2.10

Which distribution should you use for this problem?

Answer

$$N\left(244, \frac{15}{\sqrt{50}}\right)$$

? Exercise 8.2.11

Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

? Exercise 8.2.12

What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Answer

As the sample size increases, there will be less variability in the mean, so the interval size decreases.

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

? Exercise 8.2.13

Identify the following:

- $\bar{x} = \underline{\hspace{2cm}}$
- $\sigma = \underline{\hspace{2cm}}$
- $n = \underline{\hspace{2cm}}$

? Exercise 8.2.14

In words, define the random variables X and $\bar{X}$.

Answer

X is the time in minutes it takes to complete the U.S. Census short form. $\bar{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

? Exercise 8.2.15

Which distribution should you use for this problem?

? Exercise 8.2.16

Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$CI : (7.9441, 8.4559)$

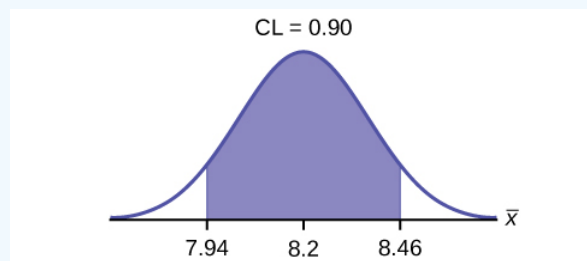


Figure 9.S. 3.

$EBM = 0.26$

? Exercise 8.2.17

If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

? Exercise 8.2.18

If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

Answer

The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

? Exercise 8.2.19

Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

? Exercise 8.2.20

Identify the following:

- a. $\bar{x}$ = _____
- b. σ = _____
- c. n = _____

Answer

- a. $\bar{x} = 2.2$
- b. $\sigma = 0.2$
- c. $n = 20$

? Exercise 8.2.21

In words, define the random variable X .

? Exercise 8.2.22

In words, define the random variable $\bar{X}$.

Answer

$\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

? Exercise 8.2.23

Which distribution should you use for this problem?

? Exercise 8.2.24

Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$$EBM = 0.07$$

$$CI : (2.1264, 2.2736)$$

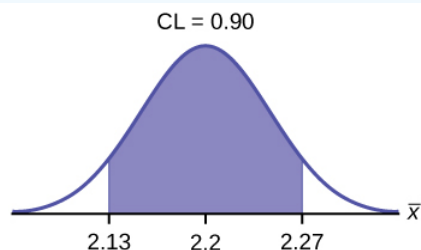


Figure 9.S. 4

? Exercise 8.2.25

Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

? Exercise 8.2.26

In complete sentences, explain why the confidence interval in Exercise is larger than in Exercise.

Answer

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

? Exercise 8.2.27

In complete sentences, give an interpretation of what the interval in Exercise means.

? Exercise 8.2.28

What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

Answer

The confidence level would increase.

? Exercise 8.2.29

What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

? Exercise 8.2.30

$\bar{x}$ = _____

Answer

30.4

? Exercise 8.2.31

n = _____

? Exercise 8.2.32

_____ = 15

Answer

σ

? Exercise 8.2.33

In words, define the random variable $\bar{X}$.

? Exercise 8.2.34

What is $\bar{x}$ estimating?

Answer

μ

? Exercise 8.2.35

Is σ_x known?

? Exercise 8.2.36

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

normal

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

? Exercise 8.2.37

How much area is in both tails (combined)? $\alpha =$ _____

? Exercise 8.2.38

How much area is in each tail? $\frac{\alpha}{2} =$ _____

Answer

0.025

? Exercise 8.2.39

Identify the following specifications:

- a. lower limit
- b. upper limit
- c. error bound

? Exercise 8.2.40

The 95% confidence interval is: _____.

Answer

(24.52, 36.28)

? Exercise 8.2.41

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

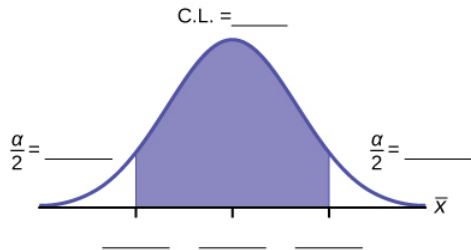


Figure 9.S. 5.

? Exercise 8.2.42

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

? Exercise 8.2.43

Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

? Exercise 8.2.44

Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Answer

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

Review

In many cases, the researcher does not know the population standard deviation, σ , of the measure being studied. In these cases, it is common to use the sample standard deviation, s , as an estimate of σ . The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t -distribution is much better. Define a t -score using the following formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (9.S.7)$$

The t -score follows the Student's t -distribution with $n-1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right) \frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the t -score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α .

Formula Review

s = the standard deviation of sample values.

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ is the formula for the t -score which measures how far away a measure is from the population mean in the Student's t -distribution

$df = n - 1$; the degrees of freedom for a Student's t -distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T , has a Student's t -distribution with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} =$ the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the t -score in the Student's t -distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

$$= (\text{point estimate} - EBM, \text{point estimate} + EBM) \quad (9.S.8)$$

$$= \left(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}} \right) \quad (9.S.9)$$

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

? Exercise 8.3.3

Identify the following:

- $\bar{x} =$ _____
- $s_x =$ _____
- $n =$ _____
- $n - 1 =$ _____

? Exercise 8.3.4

Define the random variables X and $\bar{X}$ in words.

Answer

X is the number of hours a patient waits in the emergency room before being called back to be examined. $\bar{X}$ is the mean wait time of 70 patients in the emergency room.

? Exercise 8.3.5

Which distribution should you use for this problem?

? Exercise 8.3.6

Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$CI : (1.3808, 1.6192)$

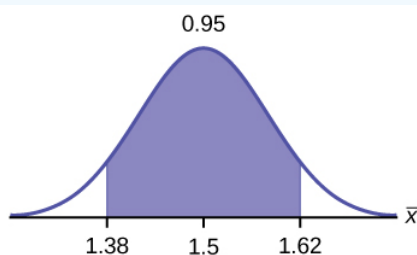


Figure 9.S. 1.

$EBM = 0.12$

? Exercise 8.3.7

Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

? Exercise 8.3.8

- $\bar{x} =$ _____
- $s_x =$ _____
- $n =$ _____
- $n - 1 =$ _____

Answer

- $\bar{x} = 151$
- $s_x = 32$
- $n = 108$
- $n - 1 = 107$

? Exercise 8.3.9

Define the random variable X in words.

? Exercise 8.3.10

Define the random variable $\bar{X}$ in words.

Answer

$\bar{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

? Exercise 8.3.11

Which distribution should you use for this problem?

? Exercise 8.3.12

Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

Answer

$CI : (142.92, 159.08)$

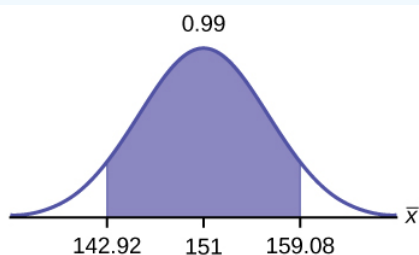


Figure 9.S. 2

$EBM = 8.08$

? Exercise 8.3.13

Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in Table are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

? Exercise 8.3.14

- a. $\bar{x}$ = _____
- b. s_x = _____
- c. n = _____

Answer

- a. 3.26
- b. 1.02
- c. 39

? Exercise 8.3.15

Define the random variable $\bar{X}$ in words.

? Exercise 8.3.16

What is $\bar{x}$ estimating?

Answer

μ

? Exercise 8.3.17

Is σ_x known?

? Exercise 8.3.18

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

t_{38}

Construct a 95% confidence interval for the true mean number of colors on national flags.

? Exercise 8.3.19

How much area is in both tails (combined)?

? Exercise 8.3.20

How much area is in each tail?

Answer

0.025

? Exercise 8.3.21

Calculate the following:

- lower limit
- upper limit
- error bound

? Exercise 8.3.22

The 95% confidence interval is_____.

Answer

(2.93, 3.59)

? Exercise 8.3.23

Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

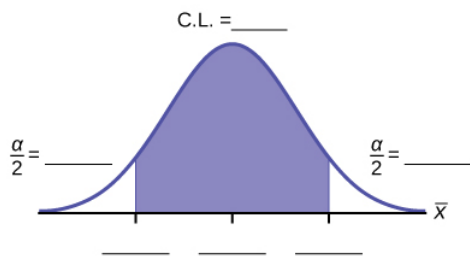


Figure 9.S. 3.

? Exercise 8.3.24

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

? Exercise 8.3.25

Using the same $\bar{x}$, s_x , and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Answer

The error bound would become $EBM = 0.245$. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

? Exercise 8.3.26

Using the same $\bar{x}$, s_x , and $n = 39$, how would the error bound change if the confidence level were reduced to 90%? Why?

References

1. Jensen, Tom. "Democrats, Republicans Divided on Opinion of Music Icons." Public Policy Polling. Available online at www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).
2. Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. "Teens, Social Media, and Privacy." PewInternet, 2013. Available online at www.pewinternet.org/Reports/2...d-Privacy.aspx (accessed July 2, 2013).
3. Prince Survey Research Associates International. "2013 Teen and Privacy Management Survey." Pew Research Center: Internet and American Life Project. Available online at www.pewinternet.org/~media/...al%20Media.pdf (accessed July 2, 2013).
4. Saad, Lydia. "Three in Four U.S. Workers Plan to Work Past Retirement Age: Slightly more say they will do this by choice rather than necessity." Gallup® Economy, 2013. Available online at <http://www.gallup.com/poll/162758/th...ement-age.aspx> (accessed July 2, 2013).
5. The Field Poll. Available online at field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).
6. Zogby. "New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure 'Investment' for National Security." Zogby Analytics, 2013. Available online at <http://www.zogbyanalytics.com/news/2...analytics-poll> (accessed July 2, 2013).
7. "52% Say Big-Time College Athletics Corrupt Education Process." Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/publ...cation_process (accessed July 2, 2013).

Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, $\frac{x}{n}$, where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$(\text{lower bound, upper bound}) = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}} \right)$$

The "plus four" method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n+4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

$p' = \frac{x}{n}$ where x represents the number of successes and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q' = 1 - p' \quad (9.S.10)$$

$$p' - N \left(p, \sqrt{\frac{pq}{n}} \right) \quad (9.S.11)$$

The variable p' has a binomial distribution that can be approximated with the normal distribution shown here.

$$EBP = \text{the error bound for a proportion} = z_{\frac{\alpha}{2}} \sqrt{\frac{p'q'}{n}}$$

Confidence interval for a proportion:

$$(\text{lower bound, upper bound}) = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}} \right)$$

$n = \frac{z_{\frac{\alpha}{2}}^2 p'q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence $1 - \alpha$ and margin of error EBP .

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$

$$EBP = \left(z_{\frac{\alpha}{2}} \right) \sqrt{\frac{p'q'}{n}} p' + q' = 1$$

The confidence interval has the format $(p' - EBP, p' + EBP)$.

- $\bar{x}$ is a point estimate for μ
- p' is a point estimate for ρ
- s is a point estimate for σ

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

? Exercise 8.4.6

When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

? Exercise 8.4.7

If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Answer

It would decrease, because the z -score would decrease, which reducing the numerator and lowering the number.

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

? Exercise 8.4.8

Identify the following:

- $x = \underline{\hspace{2cm}}$
- $n = \underline{\hspace{2cm}}$
- $p' = \underline{\hspace{2cm}}$

? Exercise 8.4.9

Define the random variables X and P' in words.

Answer

X is the number of “successes” where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

? Exercise 8.4.10

Which distribution should you use for this problem?

? Exercise 8.4.11

Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$CI : (0.5321, 0.6679)$

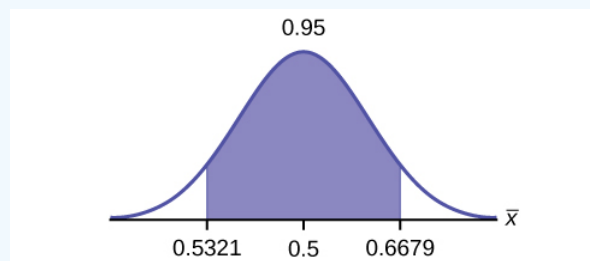


Figure 9.S. 1.

$EBM : 0.0679$

? Exercise 8.4.12

List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

? Exercise 8.4.13

We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and P' in words.

Answer

X is the number of “successes” where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

? Exercise 8.4.14

Which distribution should you use for this problem?

? Exercise 8.4.15

Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$CI : (0.19432, 0.33068)$

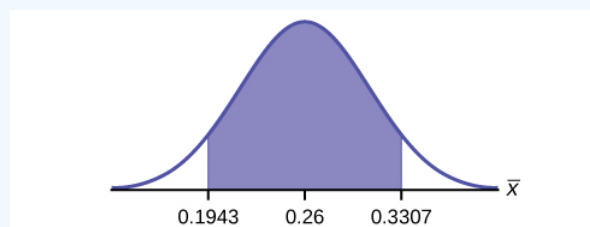


Figure 9.S. 2.

$EBM : 0.0707$

? Exercise 8.4.16

Suppose we want to lower the sampling error. What is one way to accomplish that?

? Exercise 8.4.17

The sampling error given in the survey is ± 2 . Explain what the ± 2 means.

Answer

The sampling error means that the true mean can be 2% above or below the sample mean.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

? Exercise 8.4.18

Define the random variable X in words.

? Exercise 8.4.19

Define the random variable P' in words.

Answer

P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

? Exercise 8.4.20

Which distribution should you use for this problem?

? Exercise 8.4.21

Construct a 90% confidence interval, and state the confidence interval and the error bound.

Answer

$CI : (0.62735, 0.67265)$

$EBM : 0.02265$

? Exercise 8.4.22

What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

? Exercise 8.4.23

What is being counted?

Answer

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

? Exercise 8.4.24

In words, define the random variable X .

? Exercise 8.4.25

Calculate the following:

- a. $x =$ _____
- b. $n =$ _____
- c. $p' =$ _____

Answer

- a. $x = 64$
- b. $n = 80$
- c. $p' = 0.8$

? Exercise 8.4.26

State the estimated distribution of X . $X \sim$ _____

? Exercise 8.4.27

Define a new random variable P' . What is p' estimating?

Answer

p

? Exercise 8.4.28

In words, define the random variable P' .

? Exercise 8.4.29

State the estimated distribution of P' . Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

Answer

$$P \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right) = (0.72171, 0.87829)$$

? Exercise 8.4.30

How much area is in both tails (combined)?

? Exercise 8.4.31

How much area is in each tail?

Answer

0.04

? Exercise 8.4.32

Calculate the following:

- lower limit
- upper limit
- error bound

? Exercise 8.4.33

The 92% confidence interval is _____.

Answer

(0.72; 0.88)

? Exercise 8.4.34

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

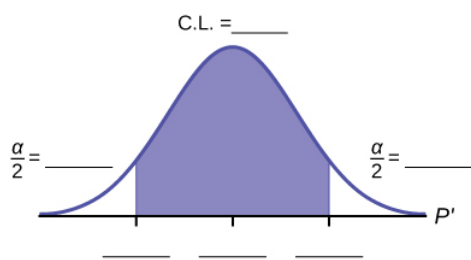


Figure 9.S. 3.

? Exercise 8.4.35

In one complete sentence, explain what the interval means.

Answer

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

? Exercise 8.4.36

Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

? Exercise 8.4.37

Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

Answer

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

? Exercise 8.4.38

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

This page titled [9.S: Confidence Intervals \(Summary\)](#) 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.

- **8.S: Confidence Intervals (Summary)** by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

CHAPTER OVERVIEW

10: Hypothesis Testing with One Sample

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of \$60,000 per year.

[10.0: Prelude to Hypothesis Testing](#)

[10.1: Null and Alternative Hypotheses](#)

[10.2: Outcomes and the Type I and Type II Errors](#)

[10.3: Distribution Needed for Hypothesis Testing](#)

[10.4: Rare Events, the Sample, Decision and Conclusion](#)

[10.5: Additional Information and Full Hypothesis Test Examples](#)

This page titled [10: Hypothesis Testing with One Sample](#) 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.

10.0: Prelude to Hypothesis Testing

CHAPTER OBJECTIVES

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

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of \$60,000 per year.



Figure 10.0.1: You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

A statistician will make a decision about these claims. This process is called "hypothesis testing." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analysis of the data, to reject the null hypothesis. In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

- Set up two contradictory hypotheses.
- Collect sample data (in homework problems, the data or summary statistics will be given to you).
- Determine the correct distribution to perform the hypothesis test.
- Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
- Make a decision and write a meaningful conclusion.

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See [Appendix E](#).

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

This page titled [10.0: Prelude to Hypothesis Testing](#) 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: Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0 : **The null hypothesis**: It is a statement of no difference between the variables—they are not related. This can often be considered the status quo and as a result if you cannot accept the null it requires some action.

H_a : **The alternative hypothesis**: It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0 . This is usually what the researcher is trying to prove.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision**. There are two options for a decision. They are "reject H_0 " if the sample information favors the alternative hypothesis or "do not reject H_0 " or "decline to reject H_0 " if the sample information is insufficient to reject the null hypothesis.

Table 10.1.1: Mathematical Symbols Used in H_0 and H_a :

H_0	H_a
equal (=)	not equal ($\neq$) or greater than ($>$) or less than ($<$)

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with $>$ or $<$ as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

✓ Example 10.1.1

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

- $H_0 : \mu = 2.0$
- $H_a : \mu \neq 2.0$

? Exercise 10.1.1

We want to test whether the mean height of eighth graders is higher than 66 inches. State the null and alternative hypotheses. Fill in the correct symbol ($=$, $\neq$, $\geq$, $<$, $\leq$, $>$) for the null and alternative hypotheses.

- $H_0 : \mu _ 66$
- $H_a : \mu _ 66$

Answer

- $H_0 : \mu = 66$
- $H_a : \mu > 66$

? Exercise 10.1.2

On a state driver's test, about 40% pass the test on the first try. We want to test if fewer than 40% pass on the first try. Fill in the correct symbol ($=$, $\neq$, $\geq$, $<$, $\leq$, $>$) for the null and alternative hypotheses.

- $H_0 : p _ 0.40$
- $H_a : p _ 0.40$

Answer

- $H_0 : p = 0.40$

b. $H_a : p < 0.40$

COLLABORATIVE EXERCISE

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0 . The null is not rejected unless the hypothesis test shows otherwise.
The null statement must always contain some form of equality ($=$, $\leq$ or $\geq$)
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1 , using less than, greater than, or not equals symbols, i.e., ($\neq$, $>$, or $<$).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

If H_a has:	equal ($=$)
then H_0 has:	not equal ($\neq$) or greater than ($>$) or less than ($<$)

- If $\alpha \leq p$ -value, then do not reject H_0 .
- If $\alpha > p$ -value, then reject H_0 .

α is preconceived. Its value is set before the hypothesis test starts. The p -value is calculated from the data. References

Data from the National Institute of Mental Health. Available online at <http://www.nimh.nih.gov/publicat/depression.cfm>.

Glossary

Hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

This page titled [10.1: Null and Alternative Hypotheses](#) 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: Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_0 is Actually True	H_0 is Actually False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject** H_0 when H_0 is **true** (**correct decision**).
2. The decision is to **reject** H_0 when H_0 is **true** (incorrect decision known as a Type I error).
3. The decision is **not to reject** H_0 when, in fact, H_0 is **false** (incorrect decision known as a Type II error).
4. The decision is to **reject** H_0 when H_0 is **false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

- α = probability of a Type I error = $P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.
- β = probability of a Type II error = $P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The *Power of the Test* is $1 - \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test. The following are examples of Type I and Type II errors.

✓ Example 10.2.1: Type I vs. Type II errors

Suppose the null hypothesis, H_0 , is: Frank's rock climbing equipment is safe.

- **Type I error:** Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe.
- **Type II error:** Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe.

β = **probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

? Exercise 10.2.1

Suppose the null hypothesis, H_0 , is: the blood cultures contain no traces of pathogen X . State the Type I and Type II errors.

Answer

- **Type I error:** The researcher thinks the blood cultures do contain traces of pathogen X , when in fact, they do not.
- **Type II error:** The researcher thinks the blood cultures do not contain traces of pathogen X , when in fact, they do.

✓ Example 10.2.2

Suppose the null hypothesis, H_0 , is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

- **Type I error:** The emergency crew thinks that the victim is dead when, in fact, the victim is alive.
- **Type II error:** The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α = **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = $P(\text{Type I error})$.

β = **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = $P(\text{Type II error})$.

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

? Exercise 10.2.2

Suppose the null hypothesis, H_0 , is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Answer

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

✓ Example 10.2.3

It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0 , is: It's a Boy Genetic Labs has no effect on gender outcome.

- **Type I error:** This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α .
- **Type II error:** This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β .

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Summary

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I error** occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected. The probabilities of these errors are denoted by the Greek letters α and β , for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

- α = probability of a Type I error = $P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.
- β = probability of a Type II error = $P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

Glossary

Type 1 Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

This page titled [10.2: Outcomes and the Type I and Type II Errors](#) 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: Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a normal distribution or a Student's t -distribution. (Remember, use a Student's t -distribution when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a single population mean, the distribution for the test is for *means*:

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \quad (10.3.1)$$

or

$$t_{df} \quad (10.3.2)$$

The population parameter is μ . The estimated value (point estimate) for μ is $\bar{x}$, the sample mean.

If you are testing a single population proportion, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{p-q}{n}}\right) \quad (10.3.3)$$

The population parameter is p . The estimated value (point estimate) for p is p' . $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean** μ using a Student's t -distribution (often called a t -test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample standard deviation to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t -test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z -test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion** p , you take a simple random sample from the population. You must meet the conditions for a binomial distribution which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p . The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than ten ($np > 10$ and $nq > 10$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

Summary

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t -test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of successes and the mean number of failures satisfy the conditions: $np > 5$ and $nq > 5$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α , then use $\alpha = 0.05$.

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's t -test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means:
 $\mu = \mu_{\bar{x}}$ and $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation.
Proportions: $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Glossary

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n , of independent trials.

“Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$ $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

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 **the standard normal distribution**.

Standard Deviation

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

Student's t -Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t -distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

This page titled [10.3: Distribution Needed for Hypothesis Testing](#) 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.

- **9.4: Distribution Needed for Hypothesis Testing** by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

10.4: Rare Events, the Sample, Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very *unlikely* to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an assumption—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a \$100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a \$100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more \$100 bills in the basket. A "rare event" has occurred (Didi getting the \$100 bill) so Ali doubts the assumption about only one \$100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the p -value. The p -value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.

A large p -value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p -value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p -value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Exercise 10.4.1

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

- $H_0 : \mu = 12$
- $H_a : \mu > 12$

The p -value is 0.0013

Draw a graph that shows the p -value.

Answer

$$p\text{-value} = 0.0013$$

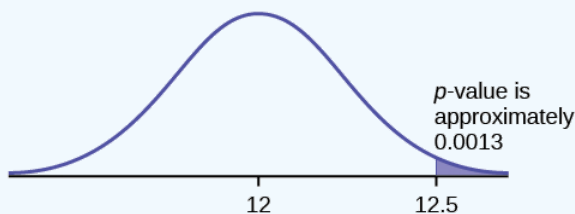


Figure 10.4.2

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p -value and a preset or preconceived α (also called a "**significance level**"). A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a decision to reject or not reject H_0 , do as follows:

- If $\alpha > p$ -value, reject H_0 . The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis, H_a , may be correct.
- If $\alpha \leq p$ -value, do not reject H_0 . The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, H_a , may be correct.

When you "do not reject H_0 ", it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0 .

Conclusion: After you make your decision, write a thoughtful conclusion about the hypotheses in terms of the given problem.

✓ Example 10.4.2

When using the p -value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

- If the p -value is low, the null must go.
- If the p -value is high, the null must fly.

This memory aid relates a p -value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates a p -value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null when hypothesis when _____.

The results of the sample data _____.

Answer

Reject the null hypothesis when **the p -value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the p -value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

? Exercise 10.4.2

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

- $H_0 : p = 0.50, H_a : p > 0.50$
- $\alpha = 0.01$
- p -value = 0.025

Interpret the results and state a conclusion in simple, non-technical terms.

Answer

Since the p -value is greater than the established alpha value (the p -value is high), we do not reject the null hypothesis. There is not enough evidence to support It's a Boy Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p -value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

- $\alpha > p - \text{value}$, reject the null hypothesis
- $\alpha \leq p - \text{value}$, do not reject the null hypothesis

Glossary

Level of Significance of the Test

probability of a Type I error (reject the null hypothesis when it is true). Notation: α . In hypothesis testing, the Level of Significance is called the preconceived α or the preset α .

p -value

the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p -value, the stronger the evidence is against the null hypothesis.

This page titled [10.4: Rare Events, the Sample, Decision and Conclusion](#) 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.

10.5: Additional Information and Full Hypothesis Test Examples

- In a hypothesis test problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α .
- The statistician setting up the hypothesis test selects the value of α to use before collecting the sample data.
- If no level of significance is given, a common standard to use is $\alpha = 0.05$.
- When you calculate the p -value and draw the picture, the p -value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The alternative hypothesis, H_a , tells you if the test is left, right, or two-tailed. It is the key to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- Thinking about the meaning of the p -value: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p -value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p -value such as 0.4, as opposed to a p -value of 0.056 ($\alpha = 0.05$ is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

✓ Example 10.5.1

$$H_0 : \mu = 5, H_a : \mu < 5$$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p -value is as follows:

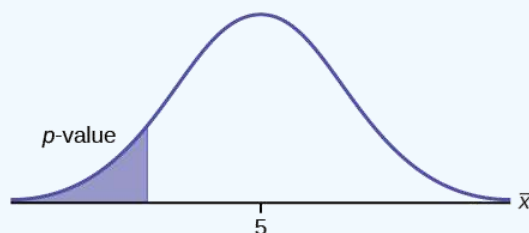


Figure 10.5.1

? Exercise 10.5.1

$$H_0 : \mu = 10, H_a : \mu < 10$$

Assume the p -value is 0.0935. What type of test is this? Draw the picture of the p -value.

Answer

left-tailed test

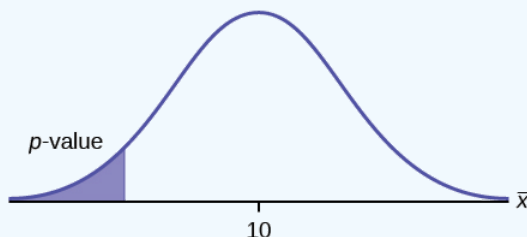


Figure 10.5.2

✓ Example 10.5.3

$$H_0 : \mu = 50, H_a : \mu \neq 50$$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p -value is as follows.

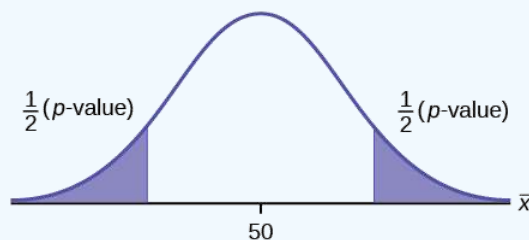


Figure 10.5.3

? Exercise 10.5.3

$$H_0 : \mu = 0.5, H_a : \mu \neq 0.5$$

Assume the p -value is 0.2564. What type of test is this? Draw the picture of the p -value.

Answer

two-tailed test

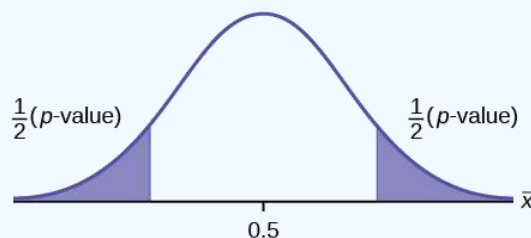


Figure 10.5.6

Full Hypothesis Test Examples

✓ Example 10.5.4

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Answer

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

$$H_0 : \mu = 65 \quad H_a : \mu > 65$$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t .

Use t_{df} . Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p -value using the Student's t -distribution:

$p\text{-value} = P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p -value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.

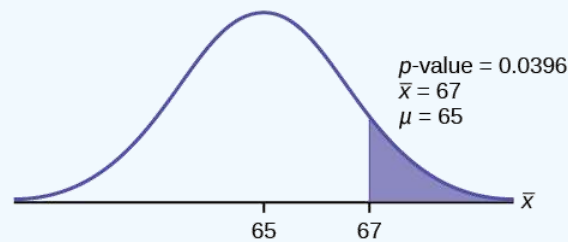


Figure 10.5.4

Compare α and the p -value:

Since $\alpha = 0.05$ and $p\text{-value} = 0.0396$, $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0 .

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p -value can easily be calculated.

Put the data into a list. Press **STAT** and arrow over to **TESTS**. Press **2:T-Test**. Arrow over to **Data** and press **ENTER**. Arrow down and enter 65 for μ_0 , the name of the list where you put the data, and 1 for **Freq:**. Arrow down to μ : and arrow over to $> \mu_0$. Press **ENTER**. Arrow down to **Calculate** and press **ENTER**. The calculator not only calculates the p -value ($p = 0.0396$) but it also calculates the test statistic (t -score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to **Draw** (instead of **Calculate**). Press **ENTER**. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p -value). Make sure when you use **Draw** that no other equations are highlighted in $Y =$ and the plots are turned off.

? Exercise 10.5.4

It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p -value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : \mu = 5$
- $H_a : \mu < 5$
- $p = 0.0082$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than \$5 a week.

- Type I Error: To conclude that the stock price is growing slower than \$5 a week when, in fact, the stock price is growing at \$5 a week (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that the stock price is growing at a rate of \$5 a week when, in fact, the stock price is growing slower than \$5 a week (do not reject the null hypothesis when the null hypothesis is false).

The remaining examples and exercises in this section deal with hypothesis testing of a population proportion.

✓ Example 10.5.5

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Answer

Set up the hypothesis test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

$$H_0 : p = 0.50 \quad H_a : p \neq 0.50$$

The words "**is the same or different from**" tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P' , the estimated proportion.

$$P' - N \left(p, \sqrt{\frac{p-q}{n}} \right)$$

Therefore,

$$P' - N \left(0.5, \sqrt{\frac{0.5 - 0.5}{100}} \right)$$

where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$

Calculate the p -value using the normal distribution for proportions:

$$p\text{-value} = P(p' < 0.47 \text{ or } p' > 0.53) = 0.5485$$

where

$$x = 53, p' = \frac{x}{n} = \frac{53}{100} = 0.53$$

Interpretation of the p -value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in Figure).

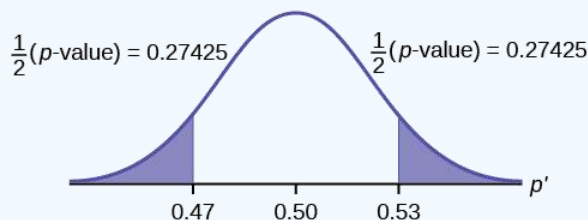


Figure 10.5.5

$\mu = p = 0.50$ comes from H_0 , the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 - 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p -value:

Since $\alpha = 0.01$ and $p\text{-value} = 0.5485$, $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, you cannot reject H_0 .

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p -value can easily be calculated.

Press **STAT** and arrow over to **TESTS**. Press **5:1-PropZTest**. Enter .5 for p_0 , 53 for x and 100 for n . Arrow down to **Prop** and arrow to **not equals** p_0 . Press **ENTER**. Arrow down to **Calculate** and press **ENTER**. The calculator calculates the p -value ($p = 0.5485$) and the test statistic (z -score). **Prop not equals .5** is the alternate hypothesis. Do this set of instructions again except arrow to **Draw** (instead of **Calculate**). Press **ENTER**. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p -value). Make sure when you use **Draw** that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

? Exercise 10.5.5

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p -value, sketch the graph, and state your conclusion.

Answer

Since the problem is about percentages, this is a test of single population proportions.

- $H_0 : p = 0.85$
- $H_a : p \neq 0.85$
- $p = 0.7554$

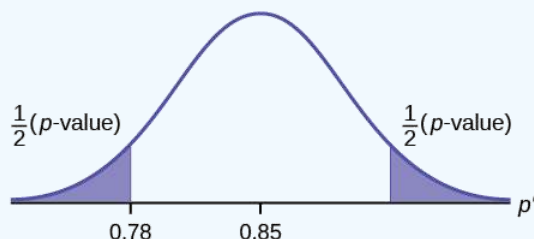


Figure 10.5.6

Because $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

✓ Example 10.5.6

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Set up the Hypothesis Test:

$$H_0 : p = 0.30, H_a : p \neq 0.30$$

Determine the distribution needed:

The **random variable** is $P' =$ proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' - N\left(0.30, \sqrt{\frac{(0.30-0.70)}{150}}\right)$

? Exercise 10.5.6.1

a. The value that helps determine the p -value is p' . Calculate p' .

Answer

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43, n = 150$

$p' = 43/150$

? Exercise 10.5.6.2

b. What is the level of significance?

Answer

b. The level of significance is the preset α . Since α is not given, assume that $\alpha = 0.05$.

? Exercise 10.5.6.3

c. Calculate the p -value.

Answer

c. p -value = 0.7216

? Exercise 10.5.6.4

d. Make a decision. _____ (Reject/Do not reject) H_0 because _____.

Answer

e. Assuming that $\alpha = 0.05, \alpha < p$ -value. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

? Exercise 10.5.7

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p -value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : p = 0.92$
- $H_a : p < 0.92$
- p -value = 0.0046

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

- Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

✓ Example 10.5.8

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Answer

We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion z-test. The hypotheses for the test will be
 - a. $H_0 : p = 0.00078$
 - b. $H_a : p \neq 0.00078$
3. The following screen shots display the summary statistics from the hypothesis test.

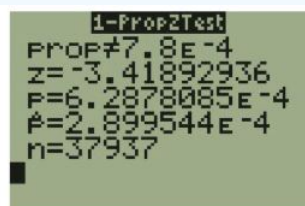


Figure 10.5.8

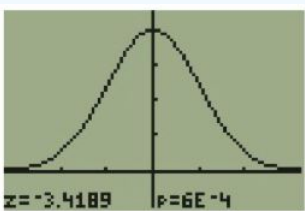


Figure 10.5.9

4. Since the p -value, $p = 0.00063$ is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.

Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a . Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p -value. (A z -score and a t -score are examples of test statistics.)
5. Compare the preconceived α with the p -value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β . β is needed to help determine the sample size of the data that is used in calculating the p -value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is desirable. If the

power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

? Exercise 10.5.9

Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

? Exercise 10.5.10

Assume $H_0 : p = 0.25$ and $H_a : p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

? Exercise 10.5.11

Draw the graph of a two-tailed test.

Answer

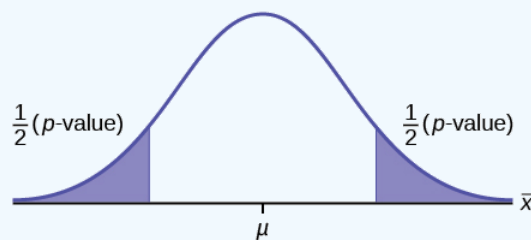


Figure 10.5.11

? Exercise 10.5.12

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

Answer

a right-tailed test

? Exercise 10.5.13

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

Answer

a left-tailed test

? Exercise 10.5.14

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at www.businessweek.com/news/2011-09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at <http://energy.gov> (accessed June 27, 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at <http://www.bls.gov/oes/current/oes291111.htm>.
10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013).
11. Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.
13. Data from Toastmasters International. Available online at <http://toastmasters.org/artisan/details?eID=429&Page=1>.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. "Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005." Available online at <http://www.disastercenter.com/kentucky/crime/3868.htm> (accessed June 27, 2013).
16. "Foothill-De Anza Community College District." De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. "Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark." *Institute of Cancer Epidemiology and the Danish Cancer Society*, 93(3):203-7. Available online at <http://www.ncbi.nlm.nih.gov/pubmed/11158188> (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. "How often does sexual assault occur?" RAINN, 2009. Available online at www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

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 distribution 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.

This page titled [10.5: Additional Information and Full Hypothesis Test Examples](#) 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.

- [9.6: Additional Information and Full Hypothesis Test Examples](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

CHAPTER OVERVIEW

11: Hypothesis Testing with Two Samples

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded. To compare two means or two proportions, you work with two groups. The groups are classified either as independent or matched pairs. Independent groups consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Matched pairs consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

[11.1: Comparing Two Independent Population Means \(Hypothesis test\)](#)

[11.2: Comparing Two Independent Population Proportions \(Hypothesis test\)](#)

[11.3: Matched or Paired Samples](#)

This page titled [11: Hypothesis Testing with Two Samples](#) 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.

11.1: Comparing Two Independent Population Means (Hypothesis test)

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations:
 - if the sample sizes are small, the distributions are important (should be normal)
 - if the sample sizes are large, the distributions are not important (need not be normal)

The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t -test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $\bar{X}_1 - \bar{X}_2$, and divide by the standard error in order to standardize the difference. The result is a t -score test statistic.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of **the difference in sample means**, $\bar{X}_1 - \bar{X}_2$.

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \quad (11.1.1)$$

The test statistic (t -score) is calculated as follows:

$$\frac{(\bar{x} - \bar{x}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \quad (11.1.2)$$

where:

- s_1 and s_2 , the sample standard deviations, are estimates of σ_1 and σ_2 , respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means. μ_1 and μ_2 are the population means.

The number of *degrees of freedom* (df) requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The df are not always a whole number. The test statistic calculated previously is approximated by the Student's t -distribution with df as follows:

Degrees of freedom

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2} \right)^2}{\left(\frac{1}{n_1 - 1} \right) \left(\frac{(s_1)^2}{n_1} \right)^2 + \left(\frac{1}{n_2 - 1} \right) \left(\frac{(s_2)^2}{n_2} \right)^2} \quad (11.1.3)$$

We can also use a conservative estimation of degree of freedom by taking DF to be the smallest of $n_1 - 1$ and $n_2 - 1$

When both sample sizes n_1 and n_2 are five or larger, the Student's t approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

It is not necessary to compute the degrees of freedom by hand. A calculator or computer easily computes it.

✓ Example 11.1.1: Independent groups

The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in Table 11.1.1. Each populations has a normal distribution.

Table 11.1.1

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

Answer

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two independent groups, two population means.

Random variable: $\bar{X}_g - \bar{X}_b =$ difference in the sample mean amount of time girls and boys play sports each day.

- $H_0 : \mu_g = \mu_b$
- $H_0 : \mu_g - \mu_b = 0$
- $H_a : \mu_g \neq \mu_b$
- $H_a : \mu_g - \mu_b \neq 0$

The words "**the same**" tell you H_0 has an "=". Since there are no other words to indicate H_a , assume it says "**is different.**" This is a two-tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8462. **Do not pool the variances.**

Calculate the p -value using a Student's t -distribution: $p\text{-value} = 0.0054$

Make a decision: Since $\alpha > p\text{-value}$, reject H_0 . This means you reject $\mu_g = \mu_b$. The means are different.

Press **STAT**. Arrow over to **TESTS** and press **4:2-SampTTest**. Arrow over to **Stats** and press **ENTER**. Arrow down and enter **2** for the first sample mean, **0.866** for $Sx1$, **9** for $n1$, **3.2** for the second sample mean, **1** for $Sx2$, and **16** for $n2$. Arrow down to $\mu1$: and arrow to **does not equal** $\mu2$. Press **ENTER**. Arrow down to **Pooled:** and **No**. Press **ENTER**. Arrow down to **Calculate** and press **ENTER**. The p -value is $p = 0.0054$, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of **Calculate** do **Draw**.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged seven to 11 play sports per day is different (mean number of hours boys aged seven to 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged seven to 11 play sports per day is greater than the mean number of hours played by boys).

? Exercise 11.1.1

Two samples are shown in Table. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5% level of significance.

Table 11.1.2

	Sample Size	Sample Mean	Sample Standard Deviation
Population A	25	5	1
Population B	16	4.7	1.2

Answer

The p -value is 0.4125 which is much higher than 0.05, so we decline to reject the null hypothesis. There is not sufficient evidence to conclude that the means of the two populations are not the same.

When both of the sample sizes is larger than 30 ($n_1 > 30$ and $n_2 > 30$) you can use the normal distribution to approximate the Student's t .

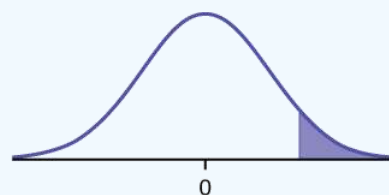
✓ Example 11.1.2

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A **has taken more math classes**, on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

- Is this a test of two means or two proportions?
- Are the populations standard deviations known or unknown?
- Which distribution do you use to perform the test?
- What is the random variable?
- What are the null and alternate hypotheses? Write the null and alternate hypotheses in words and in symbols.
- Is this test right-, left-, or two-tailed?
- What is the p -value?
- Do you reject or not reject the null hypothesis?

Answers

- two means
- unknown
- Student's t
- $\bar{X}_A - \bar{X}_B$



$$\bar{X}_A - \bar{X}_B = 0.5^*$$

$$\text{Note: } \bar{X}_A - \bar{X}_B = 4 - 3.5 = 0.5$$

- $H_0 : \mu_A = \mu_B$ and $H_a : \mu_A > \mu_B$
- right
- 0.1928
- Do not reject.
- At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

? Exercise 11.1.2

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

- Are the population standard deviations known?
- Conduct an appropriate hypothesis test. At the 5% significance level, what is your conclusion?

Answer

- They are unknown.
- The p -value = 0.0878. At the 5% level of significance, there is insufficient evidence to conclude that the workers of Company A stay longer with the company.

✓ Example 11.1.3

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in Table 11.1.3 and Table 11.1.4

Table 11.1.3: Online Class

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Table 11.1.4: Face-to-face Class

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Is the mean of the Final Exam scores of the online class lower than the mean of the Final Exam scores of the face-to-face class? Test at a 5% significance level. Answer the following questions:

- Is this a test of two means or two proportions?
- Are the population standard deviations known or unknown?
- Which distribution do you use to perform the test?
- What is the random variable?
- What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
- Is this test right, left, or two tailed?
- What is the p -value?
- Do you reject or not reject the null hypothesis?
- At the ____ level of significance, from the sample data, there ____ (is/is not) sufficient evidence to conclude that ____.

Be careful not to mix up the information for Group 1 and Group 2!

Answer

- two means
- unknown
- Student's t
- $\bar{X}_1 - \bar{X}_2$
- $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the final exam scores are equal for the online and face-to-face statistics classes.
 - $H_a : \mu_1 < \mu_2$ Alternative hypothesis: the mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.
- left-tailed
- p -value = 0.0011

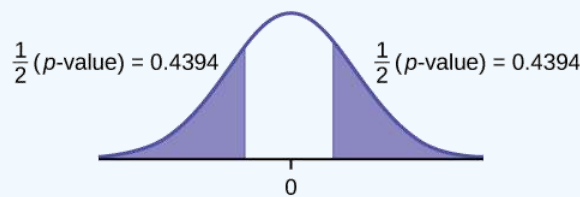


Figure 11.1.3

h. Reject the null hypothesis

i. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.

At the 5% level of significance, from the sample data, there is (is/is not) sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press 4:2SampTTest. Make sure Data is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to μ_1 : and arrow to $\neq \mu_1$ (does not equal). Press ENTER. Arrow down to Pooled: No. Press ENTER. Arrow down to Calculate and press ENTER.

✓ Example 11.1.4

Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the northeast and in the west as identified by Nasdaq on May 24, 2013 are listed in Table and Table, respectively.

Northeast

94.2	75.2	69.6	52.0	48.0	41.9	36.4	33.4	31.5	27.6
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33.0	28.7	26.5
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.5	26.0

West

126.0	70.6	65.2	51.4	45.5	37.0	33.0	29.6	23.7	22.6
116.1	70.6	58.2	51.2	43.2	36.0	31.4	28.7	23.5	21.6
78.2	68.2	55.6	50.3	39.0	34.1	31.0	25.3	23.4	21.5

Is there a difference in the weighted alpha of the top 30 stocks of banks in the northeast and in the west? Test at a 5% significance level. Answer the following questions:

- Is this a test of two means or two proportions?
- Are the population standard deviations known or unknown?
- Which distribution do you use to perform the test?
- What is the random variable?
- What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
- Is this test right, left, or two tailed?
- What is the p -value?
- Do you reject or not reject the null hypothesis?
- At the ____ level of significance, from the sample data, there ____ (is/is not) sufficient evidence to conclude that ____.

Answer

- a. two means
- b. unknown
- c. Student's-t
- d. $\bar{X}_1 - \bar{X}_2$
- e.
 - i. $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the weighted alphas are equal.
 - ii. $H_a : \mu_1 \neq \mu_2$ Alternative hypothesis : the means of the weighted alphas are not equal.
- f. two-tailed
- g. $p\text{-value} = 0.8787$
- h. Do not reject the null hypothesis (this indicates that the trends in stocks are about the same in the top 30 banks in each region)
- i. At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean weighted alphas for the banks in the northeast and the west are different

References

1. Data from Graduating Engineer + Computer Careers. Available online at www.graduatingengineer.com
2. Data from *Microsoft Bookshelf*.
3. Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).
4. "List of current United States Senators by Age." Wikipedia. Available online at en.Wikipedia.org/wiki/List_of...enators_by_age (accessed June 17, 2013).
5. "Sectoring by Industry Groups." Nasdaq. Available online at www.nasdaq.com/markets/barcha...&base=industry (accessed June 17, 2013).
6. "Strip Clubs: Where Prostitution and Trafficking Happen." Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPostrauStress.html (accessed June 17, 2013).
7. "World Series History." Baseball-Almanac, 2013. Available online at <http://www.baseball-almanac.com/ws/wsmenu.shtml> (accessed June 17, 2013).

Review

Two population means from independent samples where the population standard deviations are not known

- Random Variable: $\bar{X}_1 - \bar{X}_2 =$ the difference of the sampling means
- Distribution: Student's t -distribution with degrees of freedom (variances not pooled)

Formula Review

Standard error:

$$SE = \sqrt{\frac{(s_1^2)}{n_1} + \frac{(s_2^2)}{n_2}} \quad (11.1.4)$$

Test statistic (t -score):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \quad (11.1.5)$$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \quad (11.1.6)$$

where:

- s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.
- x_1 and x_2 are the sample means.

OR use the DF to be the smallest of $n_1 - 1$ and $n_2 - 1$

Glossary

Degrees of Freedom (df)

the number of objects in a sample that are free to vary.

Standard Deviation

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

Variable (Random Variable)

a characteristic of interest in a population being studied. Common notation for variables are upper-case Latin letters $X, Y, Z, \dots$. Common notation for a specific value from the domain (set of all possible values of a variable) are lower-case Latin letters $x, y, z, \dots$. For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if $X =$ hair color, then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the random variable X takes only after performing the experiment.

This page titled [11.1: Comparing Two Independent Population Means \(Hypothesis test\)](#) 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: Two Population Means with Unknown Standard Deviations](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/introductory-statistics>.

11.2: Comparing Two Independent Population Proportions (Hypothesis test)

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least 10, and the number of failures is at least 10, for each of the samples.
3. Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, $H_0 : p_A = p_B$. To conduct the test, we use a pooled proportion, p_c .

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \quad (11.2.1)$$

The distribution for the differences is:

$$P'_A - P'_B \sim N \left[0, \sqrt{p_c(1-p_c) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)} \right] \quad (11.2.2)$$

The test statistic (z-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c(1-p_c) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}} \quad (11.2.3)$$

✓ Example 11.2.1

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

Answer

The problem asks for a difference in proportions, making it a test of two proportions.

Let A and B be the subscripts for medication A and medication B, respectively. Then p_A and p_B are the desired population proportions.

Random Variable: $P'_A - P'_B$ = difference in the proportions of adult patients who did not react after 30 minutes to medication A and to medication B.

$$H_0 : p_A = p_B$$

$$p_A - p_B = 0$$

$$H_a : p_A \neq p_B$$

$$p_A - p_B \neq 0$$

The words "**is a difference**" tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 0.08 \quad (11.2.4)$$

$$1 - p_c = 0.92 \quad (11.2.5)$$

$$P'_A - P'_B \sim N \left[0, \sqrt{(0.08)(0.92) \left(\frac{1}{200} + \frac{1}{200} \right)} \right] \quad (11.2.6)$$

$P'_A - P'_B$ follows an approximate normal distribution.

Calculate the p -value using the normal distribution: $p\text{-value} = 0.1404$.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

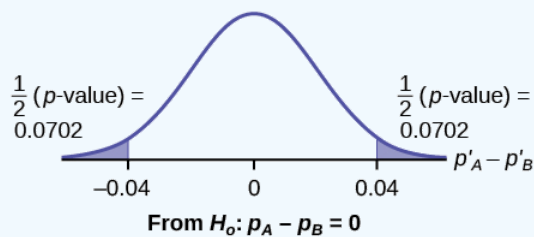


Figure 10.4.1.

$P'_A - P'_B = 0.1 - 0.06 = 0.04$.

Half the p -value is below -0.04 , and half is above 0.04 .

Compare α and the p -value: $\alpha = 0.01$ and the $p\text{-value} = 0.1404$. $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0 .

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

Press **STAT**. Arrow over to **TESTS** and press **6:2-PropZTest**. Arrow down and enter **20** for **x1**, **200** for **n1**, **12** for **x2**, and **200** for **n2**. Arrow down to **p1**: and arrow to **not equal p2**. Press **ENTER**. Arrow down to **Calculate** and press **ENTER**. The p -value is $p = 0.1404$ and the test statistic is **1.47**. Do the procedure again, but instead of **Calculate** do **Draw**.

? Exercise 11.2.1

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve A cracked under 4,500 psi. Six out of a random sample of 100 of Valve B cracked under 4,500 psi. Test at a 5% level of significance.

Answer

The p -value is 0.0379, so we can reject the null hypothesis. At the 5% significance level, the data support that there is a difference in the pressure tolerances between the two valves.

✓ Example 11.2.2: Sexting

A research study was conducted about gender differences in “sexting.” The researcher believed that the proportion of girls involved in “sexting” is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in Table. Is the proportion of girls sending sexts less than the proportion of boys “sexting?” Test at a 1% level of significance.

Males	Females

	Males	Females
Sent "sexts"	183	156
Total number surveyed	2231	2169

Answer

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then p_M and p_F are the desired population proportions.

Random variable: $p'_F - p'_M$ = difference in the proportions of males and females who sent "sexts."

$$H_0 : p_F = p_m \quad H_0 : p_F - p_M = 0$$

$$H_a : p_F < p_m \quad H_a : p_F - p_M < 0$$

The words "less than" tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_C = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077 \quad (11.2.7)$$

$$1 - p_C = 0.923 \quad (11.2.8)$$

Therefore,

$$p'_F - p'_M \sim N \left(0, \sqrt{(0.077)(0.923) \left(\frac{1}{2169} + \frac{1}{2231} \right)} \right) \quad (11.2.9)$$

$p'_F - p'_M$ follows an approximate normal distribution.

Calculate the p -value using the normal distribution:

$$p\text{-value} = 0.1045$$

Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082

Graph:

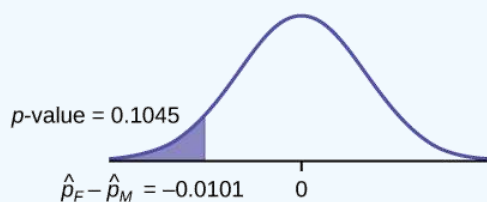


Figure 10.4.2.

Decision: Since $\alpha < p\text{-value}$, Do not reject H_0

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending "sexts" is less than the proportion of boys sending "sexts."

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to p1: and arrow to less than p2. Press ENTER. Arrow down to Calculate and press ENTER. The p -value is $P = 0.1045$ and the test statistic is $z = -1.256$.

✓ Example 11.2.3

Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232

African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of white iPhone owners greater than the proportion of African American iPhone owners?

Answer

This is a test of two population proportions. Let W and A be the subscripts for the whites and African Americans. Then p_W and p_A are the desired population proportions.

Random variable: $p'_W - p'_A$ = difference in the proportions of Android and iPhone users.

$$H_0 : p_W = p_A \quad H_0 : p_W - p_A = 0$$

$$H_a : p_W > p_A \quad H_a : p_W - p_A > 0$$

The words "more popular" indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal:

$$p_C = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927 \quad (11.2.10)$$

$$1 - p_C = 0.9073 \quad (11.2.11)$$

Therefore,

$$p'_W - p'_A \sim N \left(0, \sqrt{(0.0927)(0.9073) \left(\frac{1}{1343} + \frac{1}{232} \right)} \right) \quad (11.2.12)$$

$p'_W - p'_A$ follows an approximate normal distribution.

Calculate the p -value using the normal distribution:

$$p\text{-value} = 0.0077$$

Estimated proportion for group A: 0.10

Estimated proportion for group B: 0.05

Graph:

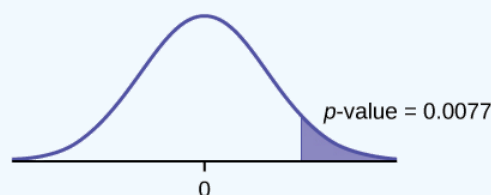


Figure 10.4.3.

Decision: Since $\alpha > p\text{-value}$, reject the H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

TI-83+ and TI-84: Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 135 for x1, 1343 for n1, 12 for x2, and 232 for n2. Arrow down to p1: and arrow to greater than p2. Press ENTER. Arrow down to Calculate and press ENTER. The P-value is $P = 0.0092$ and the test statistic is $Z = 2.33$.

? Exercise 11.2.3

A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category. Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?

- Which distribution do you use to perform the test?
- What is the random variable?
- What are the null and alternative hypothesis? Write the null and alternative hypothesis in symbols.
- Is this test right-, left-, or two-tailed?
- What is the p -value?
- Do you reject or not reject the null hypothesis?
- At the ____ level of significance, from the sample data, there ____ (is/is not) sufficient evidence to conclude that ____.

Answers

- two proportions
- normal for two proportions
- Subscripts: 1 = 2010, 2 = 2011 $P'_1 - P'_2$
- $H_0 : p_1 = p_2$
 $H_1 : p_1 \neq p_2$
- two-tailed
- p -value = 0.00086

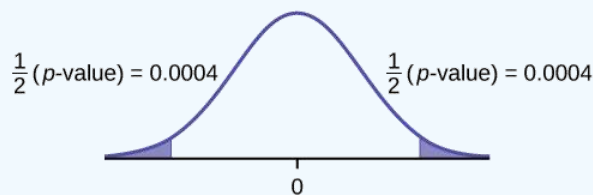


Figure 10.4.4.

- Reject the H_0 .
- At the 5% significance level, from the sample data, there is sufficient evidence to conclude that there is a difference between the proportion of forcible rapes in 2011 and 2010.

References

- Data from *Educational Resources*, December catalog.
- Data from Hilton Hotels. Available online at <http://www.hilton.com> (accessed June 17, 2013).
- Data from Hyatt Hotels. Available online at hyatt.com (accessed June 17, 2013).
- Data from Statistics, United States Department of Health and Human Services.
- Data from Whitney Exhibit on loan to San Jose Museum of Art.
- Data from the American Cancer Society. Available online at <http://www.cancer.org/index> (accessed June 17, 2013).
- Data from the Chancellor's Office, California Community Colleges, November 1994.
- "State of the States." Gallup, 2013. Available online at www.gallup.com/poll/125066/St...ef=interactive (accessed June 17, 2013).
- "West Nile Virus." Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/ncidod/dvbid/westnile/index.htm> (accessed June 17, 2013).

Review

- Test of two population proportions from independent samples.
- Random variable: $\hat{p}_A - \hat{p}_B$ = difference between the two estimated proportions
- Distribution: normal distribution

Formula Review

Pooled Proportion:

$$p_c = \frac{x_F + x_M}{n_F + n_M} \quad (11.2.13)$$

Distribution for the differences:

$$p'_A - p'_B \sim N \left[0, \sqrt{p_c(1-p_c) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)} \right] \quad (11.2.14)$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

Test Statistic (z-score):

$$z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}} \quad (11.2.15)$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

and

- p'_A and p'_B are the sample proportions, p_A and p_B are the population proportions,
- P_c is the pooled proportion, and n_A and n_B are the sample sizes.

Glossary

Pooled Proportion

estimate of the common value of p_1 and p_2 .

This page titled [11.2: Comparing Two Independent Population Proportions \(Hypothesis test\)](#) 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.

11.3: Matched or Paired Samples

This page titled [11.3: Matched or Paired Samples](#) 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.

Index

B

bar graph

2.1: Line Graphs and Bar Graphs

Bernoulli trial

5.3: Binomial Distribution

binomial probability distribution

5.3: Binomial Distribution

9.3: A Population Proportion

blinding

1.4: Experimental Design and Ethics

box plots

2.4: Box Plots

C

central limit theorem

8.3: Using the Central Limit Theorem

central limit theorem for sums

8.2: The Central Limit Theorem for Sums

Cluster Sampling

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

coefficient of determination

3.3: The Regression Equation

Cohen's Standards

11.1: Comparing Two Independent Population Means (Hypothesis test)

Comparing Two Population Proportions

11.2: Comparing Two Independent Population Proportions (Hypothesis test)

complement

4.1: Terminology

conditional probability

4.1: Terminology

confidence interval

10.0: Prelude to Hypothesis Testing

continuous data

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

control group

1.4: Experimental Design and Ethics

cumulative probability distributions

6.0: Introduction

cumulative relative frequency

1.3: Frequency and Frequency Tables

D

Decision

10.4: Rare Events, the Sample, Decision and Conclusion

degrees of freedom

11.1: Comparing Two Independent Population Means (Hypothesis test)

direction of a relationship between the variables

3.2: Scatter Plots

discrete data

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

Distribution for the differences

11.2: Comparing Two Independent Population Proportions (Hypothesis test)

E

ethics

1.4: Experimental Design and Ethics

event

4.1: Terminology

expected value

5.2: Mean and Standard Deviation

experimental unit

1.4: Experimental Design and Ethics

explanatory variable

1.4: Experimental Design and Ethics

extrapolation

3.4: Prediction

F

frequency

1.3: Frequency and Frequency Tables

Frequency Polygons

2.2: Histograms and Time Series Graphs

frequency table

1.3: Frequency and Frequency Tables

H

Histograms

2.2: Histograms and Time Series Graphs

hypothesis testing

10.0: Prelude to Hypothesis Testing

10.1: Null and Alternative Hypotheses

10.3: Distribution Needed for Hypothesis Testing

10.5: Additional Information and Full Hypothesis Test Examples

I

inferential statistics

9.0: Prelude to Confidence Intervals

Institutional Review Board

1.4: Experimental Design and Ethics

interpolation

3.4: Prediction

L

Law of Large Numbers

8.3: Using the Central Limit Theorem

level of measurement

1.3: Frequency and Frequency Tables

Line Graph

2.1: Line Graphs and Bar Graphs

linear correlation coefficient

3.3: The Regression Equation

linear equations

3.1: Linear Equations

LINEAR REGRESSION MODEL

3.3: The Regression Equation

lurking variable

1.4: Experimental Design and Ethics

M

margin of error

9.1: A Single Population Mean using the Normal Distribution

Matched Samples

11.3: Matched or Paired Samples

mean

2.6: Skewness and the Mean, Median, and Mode

5.2: Mean and Standard Deviation

mean for sums

8.2: The Central Limit Theorem for Sums

median

2.3: Measures of the Location of the Data

2.5: Measures of the Center of the Data

2.6: Skewness and the Mean, Median, and Mode

mode

2.5: Measures of the Center of the Data

2.6: Skewness and the Mean, Median, and Mode

N

Normal Approximation to the Binomial Distribution

8.3: Using the Central Limit Theorem

normal distribution

7.2: Using the Normal Distribution

8.1: The Central Limit Theorem for Sample Means (Averages)

O

outcome

4.1: Terminology

Outliers

2.3: Measures of the Location of the Data

P

Paired Samples

11.3: Matched or Paired Samples

parameter

1.1: Definitions of Statistics, Probability, and Key Terms

Pareto chart

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

placebo

1.4: Experimental Design and Ethics

Pooled Proportion

11.2: Comparing Two Independent Population Proportions (Hypothesis test)

population

1.1: Definitions of Statistics, Probability, and Key Terms

population mean

2.5: Measures of the Center of the Data

Population Standard Deviation

2.7: Measures of the Spread of the Data

power of the test

10.2: Outcomes and the Type I and Type II Errors

10.5: Additional Information and Full Hypothesis Test Examples

prediction

3.4: Prediction

probability

1.1: Definitions of Statistics, Probability, and Key Terms

probability distribution function

5.1: Probability Distribution Function (PDF) for a Discrete Random Variable

7.2: Using the Normal Distribution

Q

qualitative data

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

quantitative data

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

quartiles

[2.3: Measures of the Location of the Data](#)

R

random assignment

[1.4: Experimental Design and Ethics](#)

rare events

[10.4: Rare Events, the Sample, Decision and Conclusion](#)

response variable

[1.4: Experimental Design and Ethics](#)

rounding

[1.3: Frequency and Frequency Tables](#)

S

sample mean

[2.5: Measures of the Center of the Data](#)

sample space

[4.1: Terminology](#)

sample Standard Deviation

[2.7: Measures of the Spread of the Data](#)

sampling

[1: Sampling and Data](#)

Sampling Bias

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

sampling distribution of the mean

[8.1: The Central Limit Theorem for Sample Means \(Averages\)](#)

Sampling Error

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

Sampling with Replacement

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

Sampling without Replacement

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

scatter plot

[3.2: Scatter Plots](#)

significance level

[10.4: Rare Events, the Sample, Decision and Conclusion](#)

Skewed

[2.4: Box Plots](#)

[2.6: Skewness and the Mean, Median, and Mode](#)

slope

[3.1: Linear Equations](#)

Standard Deviation

[2.7: Measures of the Spread of the Data](#)

[5.2: Mean and Standard Deviation](#)

Standard deviation for Sums

[8.2: The Central Limit Theorem for Sums](#)

standard error

[11.1: Comparing Two Independent Population Means \(Hypothesis test\)](#)

Standard Error of the Mean

[8.1: The Central Limit Theorem for Sample Means \(Averages\)](#)

standard normal distribution

[7.0: Prelude to The Normal Distribution](#)

[7.1: The Standard Normal Distribution](#)

statistic

[1.1: Definitions of Statistics, Probability, and Key Terms](#)

stemplot

[2.1: Line Graphs and Bar Graphs](#)

strength of a relationship between the variables

[3.2: Scatter Plots](#)

T

test statistic

[11.3: Matched or Paired Samples](#)

The alternative hypothesis

[10.1: Null and Alternative Hypotheses](#)

The AND Event

[4.1: Terminology](#)

the central limit theorem

[8: The Central Limit Theorem](#)

The null hypothesis

[10.1: Null and Alternative Hypotheses](#)

The Or Event

[4.1: Terminology](#)

Time Series Graphs

[2.2: Histograms and Time Series Graphs](#)

treatments

[1.4: Experimental Design and Ethics](#)

type I error

[10.2: Outcomes and the Type I and Type II Errors](#)

type II error

[10.2: Outcomes and the Type I and Type II Errors](#)

U

Uniform distribution

[6.2: The Uniform Distribution](#)

V

variable

[1.1: Definitions of Statistics, Probability, and Key Terms](#)

Glossary

Detailed Licensing

Overview

Title: Math 10: Elementary Statistics (Hwang)

Webpages: 72

All licenses found:

- [CC BY 4.0](#): 87.5% (63 pages)
- [Undeclared](#): 12.5% (9 pages)

By Page

- [Math 10: Elementary Statistics \(Hwang\)](#) - *Undeclared*
 - [Front Matter](#) - *CC BY 4.0*
 - [Licensing](#) - *Undeclared*
 - [TitlePage](#) - *Undeclared*
 - [InfoPage](#) - *Undeclared*
 - [Table of Contents](#) - *Undeclared*
 - [1: Sampling and Data](#) - *CC BY 4.0*
 - [1.0: Introduction](#) - *CC BY 4.0*
 - [1.1: Definitions of Statistics, Probability, and Key Terms](#) - *CC BY 4.0*
 - [1.2: Data, Sampling, and Variation in Data and Sampling](#) - *CC BY 4.0*
 - [1.3: Frequency and Frequency Tables](#) - *CC BY 4.0*
 - [1.4: Experimental Design and Ethics](#) - *CC BY 4.0*
 - [2: Descriptive Statistics](#) - *CC BY 4.0*
 - [2.0: Prelude to Descriptive Statistics](#) - *CC BY 4.0*
 - [2.1: Line Graphs and Bar Graphs](#) - *CC BY 4.0*
 - [2.2: Histograms and Time Series Graphs](#) - *CC BY 4.0*
 - [2.3: Measures of the Location of the Data](#) - *CC BY 4.0*
 - [2.4: Box Plots](#) - *CC BY 4.0*
 - [2.5: Measures of the Center of the Data](#) - *CC BY 4.0*
 - [2.6: Skewness and the Mean, Median, and Mode](#) - *CC BY 4.0*
 - [2.7: Measures of the Spread of the Data](#) - *CC BY 4.0*
 - [3: Linear Regression and Correlation](#) - *CC BY 4.0*
 - [3.0: Prelude to Linear Regression and Correlation](#) - *CC BY 4.0*
 - [3.1: Linear Equations](#) - *CC BY 4.0*
 - [3.2: Scatter Plots](#) - *CC BY 4.0*
 - [3.3: The Regression Equation](#) - *CC BY 4.0*
 - [3.4: Prediction](#) - *CC BY 4.0*
 - [4: Probability Topics](#) - *CC BY 4.0*
 - [4.0: Introduction](#) - *CC BY 4.0*
 - [4.1: Terminology](#) - *CC BY 4.0*
 - [5: Discrete Random Variables](#) - *CC BY 4.0*
 - [5.0: Prelude to Discrete Random Variables](#) - *CC BY 4.0*
 - [5.1: Probability Distribution Function \(PDF\) for a Discrete Random Variable](#) - *CC BY 4.0*
 - [5.2: Mean and Standard Deviation](#) - *CC BY 4.0*
 - [5.3: Binomial Distribution](#) - *CC BY 4.0*
 - [6: Continuous Random Variables](#) - *CC BY 4.0*
 - [6.0: Introduction](#) - *CC BY 4.0*
 - [6.1: Continuous Probability Functions](#) - *CC BY 4.0*
 - [6.2: The Uniform Distribution](#) - *CC BY 4.0*
 - [7: The Normal Distribution](#) - *CC BY 4.0*
 - [7.0: Prelude to The Normal Distribution](#) - *CC BY 4.0*
 - [7.1: The Standard Normal Distribution](#) - *CC BY 4.0*
 - [7.2: Using the Normal Distribution](#) - *CC BY 4.0*
 - [8: The Central Limit Theorem](#) - *CC BY 4.0*
 - [8.0: Prelude to the Central Limit Theorem](#) - *CC BY 4.0*
 - [8.1: The Central Limit Theorem for Sample Means \(Averages\)](#) - *CC BY 4.0*
 - [8.2: The Central Limit Theorem for Sums](#) - *CC BY 4.0*
 - [8.3: Using the Central Limit Theorem](#) - *CC BY 4.0*
 - [9: Confidence Intervals](#) - *CC BY 4.0*
 - [9.0: Prelude to Confidence Intervals](#) - *CC BY 4.0*
 - [9.1: A Single Population Mean using the Normal Distribution](#) - *CC BY 4.0*
 - [9.2: A Single Population Mean using the Student t-Distribution](#) - *CC BY 4.0*
 - [9.3: A Population Proportion](#) - *CC BY 4.0*
 - [9.S: Confidence Intervals \(Summary\)](#) - *CC BY 4.0*
 - [10: Hypothesis Testing with One Sample](#) - *CC BY 4.0*
 - [10.0: Prelude to Hypothesis Testing](#) - *CC BY 4.0*
 - [10.1: Null and Alternative Hypotheses](#) - *CC BY 4.0*
 - [10.2: Outcomes and the Type I and Type II Errors](#) - *CC BY 4.0*
 - [10.3: Distribution Needed for Hypothesis Testing](#) - *CC BY 4.0*

- 10.4: Rare Events, the Sample, Decision and Conclusion - *CC BY 4.0*
- 10.5: Additional Information and Full Hypothesis Test Examples - *CC BY 4.0*
- 11: Hypothesis Testing with Two Samples - *CC BY 4.0*
 - 11.1: Comparing Two Independent Population Means (Hypothesis test) - *CC BY 4.0*
 - 11.2: Comparing Two Independent Population Proportions (Hypothesis test) - *CC BY 4.0*
 - 11.3: Matched or Paired Samples - *CC BY 4.0*
- Back Matter - *CC BY 4.0*
 - Index - *CC BY 4.0*
 - Glossary - *CC BY 4.0*
 - Detailed Licensing - *Undeclared*
 - Index - *Undeclared*
 - Glossary - *Undeclared*
 - Detailed Licensing - *Undeclared*

Index

B

bar graph

2.1: Line Graphs and Bar Graphs

Bernoulli trial

5.3: Binomial Distribution

binomial probability distribution

5.3: Binomial Distribution

9.3: A Population Proportion

blinding

1.4: Experimental Design and Ethics

box plots

2.4: Box Plots

C

central limit theorem

8.3: Using the Central Limit Theorem

central limit theorem for sums

8.2: The Central Limit Theorem for Sums

Cluster Sampling

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

coefficient of determination

3.3: The Regression Equation

Cohen's Standards

11.1: Comparing Two Independent Population Means (Hypothesis test)

Comparing Two Population Proportions

11.2: Comparing Two Independent Population Proportions (Hypothesis test)

complement

4.1: Terminology

conditional probability

4.1: Terminology

confidence interval

10.0: Prelude to Hypothesis Testing

continuous data

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

control group

1.4: Experimental Design and Ethics

cumulative probability distributions

6.0: Introduction

cumulative relative frequency

1.3: Frequency and Frequency Tables

D

Decision

10.4: Rare Events, the Sample, Decision and Conclusion

degrees of freedom

11.1: Comparing Two Independent Population Means (Hypothesis test)

direction of a relationship between the variables

3.2: Scatter Plots

discrete data

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

Distribution for the differences

11.2: Comparing Two Independent Population Proportions (Hypothesis test)

E

ethics

1.4: Experimental Design and Ethics

event

4.1: Terminology

expected value

5.2: Mean and Standard Deviation

experimental unit

1.4: Experimental Design and Ethics

explanatory variable

1.4: Experimental Design and Ethics

extrapolation

3.4: Prediction

F

frequency

1.3: Frequency and Frequency Tables

Frequency Polygons

2.2: Histograms and Time Series Graphs

frequency table

1.3: Frequency and Frequency Tables

H

Histograms

2.2: Histograms and Time Series Graphs

hypothesis testing

10.0: Prelude to Hypothesis Testing

10.1: Null and Alternative Hypotheses

10.3: Distribution Needed for Hypothesis Testing

10.5: Additional Information and Full Hypothesis Test Examples

I

inferential statistics

9.0: Prelude to Confidence Intervals

Institutional Review Board

1.4: Experimental Design and Ethics

interpolation

3.4: Prediction

L

Law of Large Numbers

8.3: Using the Central Limit Theorem

level of measurement

1.3: Frequency and Frequency Tables

Line Graph

2.1: Line Graphs and Bar Graphs

linear correlation coefficient

3.3: The Regression Equation

linear equations

3.1: Linear Equations

LINEAR REGRESSION MODEL

3.3: The Regression Equation

lurking variable

1.4: Experimental Design and Ethics

M

margin of error

9.1: A Single Population Mean using the Normal Distribution

Matched Samples

11.3: Matched or Paired Samples

mean

2.6: Skewness and the Mean, Median, and Mode

5.2: Mean and Standard Deviation

mean for sums

8.2: The Central Limit Theorem for Sums

median

2.3: Measures of the Location of the Data

2.5: Measures of the Center of the Data

2.6: Skewness and the Mean, Median, and Mode

mode

2.5: Measures of the Center of the Data

2.6: Skewness and the Mean, Median, and Mode

N

Normal Approximation to the Binomial Distribution

8.3: Using the Central Limit Theorem

normal distribution

7.2: Using the Normal Distribution

8.1: The Central Limit Theorem for Sample Means (Averages)

O

outcome

4.1: Terminology

Outliers

2.3: Measures of the Location of the Data

P

Paired Samples

11.3: Matched or Paired Samples

parameter

1.1: Definitions of Statistics, Probability, and Key Terms

Pareto chart

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

placebo

1.4: Experimental Design and Ethics

Pooled Proportion

11.2: Comparing Two Independent Population Proportions (Hypothesis test)

population

1.1: Definitions of Statistics, Probability, and Key Terms

population mean

2.5: Measures of the Center of the Data

Population Standard Deviation

2.7: Measures of the Spread of the Data

power of the test

10.2: Outcomes and the Type I and Type II Errors

10.5: Additional Information and Full Hypothesis Test Examples

prediction

3.4: Prediction

probability

1.1: Definitions of Statistics, Probability, and Key Terms

probability distribution function

5.1: Probability Distribution Function (PDF) for a Discrete Random Variable

7.2: Using the Normal Distribution

Q

qualitative data

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

quantitative data

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

quartiles

[2.3: Measures of the Location of the Data](#)

R

random assignment

[1.4: Experimental Design and Ethics](#)

rare events

[10.4: Rare Events, the Sample, Decision and Conclusion](#)

response variable

[1.4: Experimental Design and Ethics](#)

rounding

[1.3: Frequency and Frequency Tables](#)

S

sample mean

[2.5: Measures of the Center of the Data](#)

sample space

[4.1: Terminology](#)

sample Standard Deviation

[2.7: Measures of the Spread of the Data](#)

sampling

[1: Sampling and Data](#)

Sampling Bias

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

sampling distribution of the mean

[8.1: The Central Limit Theorem for Sample Means \(Averages\)](#)

Sampling Error

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

Sampling with Replacement

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

Sampling without Replacement

[1.2: Data, Sampling, and Variation in Data and Sampling](#)

scatter plot

[3.2: Scatter Plots](#)

significance level

[10.4: Rare Events, the Sample, Decision and Conclusion](#)

Skewed

[2.4: Box Plots](#)

[2.6: Skewness and the Mean, Median, and Mode](#)

slope

[3.1: Linear Equations](#)

Standard Deviation

[2.7: Measures of the Spread of the Data](#)

[5.2: Mean and Standard Deviation](#)

Standard deviation for Sums

[8.2: The Central Limit Theorem for Sums](#)

standard error

[11.1: Comparing Two Independent Population Means \(Hypothesis test\)](#)

Standard Error of the Mean

[8.1: The Central Limit Theorem for Sample Means \(Averages\)](#)

standard normal distribution

[7.0: Prelude to The Normal Distribution](#)

[7.1: The Standard Normal Distribution](#)

statistic

[1.1: Definitions of Statistics, Probability, and Key Terms](#)

stemplot

[2.1: Line Graphs and Bar Graphs](#)

strength of a relationship between the variables

[3.2: Scatter Plots](#)

T

test statistic

[11.3: Matched or Paired Samples](#)

The alternative hypothesis

[10.1: Null and Alternative Hypotheses](#)

The AND Event

[4.1: Terminology](#)

the central limit theorem

[8: The Central Limit Theorem](#)

The null hypothesis

[10.1: Null and Alternative Hypotheses](#)

The Or Event

[4.1: Terminology](#)

Time Series Graphs

[2.2: Histograms and Time Series Graphs](#)

treatments

[1.4: Experimental Design and Ethics](#)

type I error

[10.2: Outcomes and the Type I and Type II Errors](#)

type II error

[10.2: Outcomes and the Type I and Type II Errors](#)

U

Uniform distribution

[6.2: The Uniform Distribution](#)

V

variable

[1.1: Definitions of Statistics, Probability, and Key Terms](#)

Detailed Licensing

Overview

Title: Math 10: Elementary Statistics (Hwang)

Webpages: 72

All licenses found:

- [CC BY 4.0](#): 87.5% (63 pages)
- [Undeclared](#): 12.5% (9 pages)

By Page

- [Math 10: Elementary Statistics \(Hwang\)](#) - *Undeclared*
 - [Front Matter](#) - *CC BY 4.0*
 - [Licensing](#) - *Undeclared*
 - [TitlePage](#) - *Undeclared*
 - [InfoPage](#) - *Undeclared*
 - [Table of Contents](#) - *Undeclared*
 - [1: Sampling and Data](#) - *CC BY 4.0*
 - [1.0: Introduction](#) - *CC BY 4.0*
 - [1.1: Definitions of Statistics, Probability, and Key Terms](#) - *CC BY 4.0*
 - [1.2: Data, Sampling, and Variation in Data and Sampling](#) - *CC BY 4.0*
 - [1.3: Frequency and Frequency Tables](#) - *CC BY 4.0*
 - [1.4: Experimental Design and Ethics](#) - *CC BY 4.0*
 - [2: Descriptive Statistics](#) - *CC BY 4.0*
 - [2.0: Prelude to Descriptive Statistics](#) - *CC BY 4.0*
 - [2.1: Line Graphs and Bar Graphs](#) - *CC BY 4.0*
 - [2.2: Histograms and Time Series Graphs](#) - *CC BY 4.0*
 - [2.3: Measures of the Location of the Data](#) - *CC BY 4.0*
 - [2.4: Box Plots](#) - *CC BY 4.0*
 - [2.5: Measures of the Center of the Data](#) - *CC BY 4.0*
 - [2.6: Skewness and the Mean, Median, and Mode](#) - *CC BY 4.0*
 - [2.7: Measures of the Spread of the Data](#) - *CC BY 4.0*
 - [3: Linear Regression and Correlation](#) - *CC BY 4.0*
 - [3.0: Prelude to Linear Regression and Correlation](#) - *CC BY 4.0*
 - [3.1: Linear Equations](#) - *CC BY 4.0*
 - [3.2: Scatter Plots](#) - *CC BY 4.0*
 - [3.3: The Regression Equation](#) - *CC BY 4.0*
 - [3.4: Prediction](#) - *CC BY 4.0*
 - [4: Probability Topics](#) - *CC BY 4.0*
 - [4.0: Introduction](#) - *CC BY 4.0*
 - [4.1: Terminology](#) - *CC BY 4.0*
 - [5: Discrete Random Variables](#) - *CC BY 4.0*
 - [5.0: Prelude to Discrete Random Variables](#) - *CC BY 4.0*
 - [5.1: Probability Distribution Function \(PDF\) for a Discrete Random Variable](#) - *CC BY 4.0*
 - [5.2: Mean and Standard Deviation](#) - *CC BY 4.0*
 - [5.3: Binomial Distribution](#) - *CC BY 4.0*
 - [6: Continuous Random Variables](#) - *CC BY 4.0*
 - [6.0: Introduction](#) - *CC BY 4.0*
 - [6.1: Continuous Probability Functions](#) - *CC BY 4.0*
 - [6.2: The Uniform Distribution](#) - *CC BY 4.0*
 - [7: The Normal Distribution](#) - *CC BY 4.0*
 - [7.0: Prelude to The Normal Distribution](#) - *CC BY 4.0*
 - [7.1: The Standard Normal Distribution](#) - *CC BY 4.0*
 - [7.2: Using the Normal Distribution](#) - *CC BY 4.0*
 - [8: The Central Limit Theorem](#) - *CC BY 4.0*
 - [8.0: Prelude to the Central Limit Theorem](#) - *CC BY 4.0*
 - [8.1: The Central Limit Theorem for Sample Means \(Averages\)](#) - *CC BY 4.0*
 - [8.2: The Central Limit Theorem for Sums](#) - *CC BY 4.0*
 - [8.3: Using the Central Limit Theorem](#) - *CC BY 4.0*
 - [9: Confidence Intervals](#) - *CC BY 4.0*
 - [9.0: Prelude to Confidence Intervals](#) - *CC BY 4.0*
 - [9.1: A Single Population Mean using the Normal Distribution](#) - *CC BY 4.0*
 - [9.2: A Single Population Mean using the Student t-Distribution](#) - *CC BY 4.0*
 - [9.3: A Population Proportion](#) - *CC BY 4.0*
 - [9.S: Confidence Intervals \(Summary\)](#) - *CC BY 4.0*
 - [10: Hypothesis Testing with One Sample](#) - *CC BY 4.0*
 - [10.0: Prelude to Hypothesis Testing](#) - *CC BY 4.0*
 - [10.1: Null and Alternative Hypotheses](#) - *CC BY 4.0*
 - [10.2: Outcomes and the Type I and Type II Errors](#) - *CC BY 4.0*
 - [10.3: Distribution Needed for Hypothesis Testing](#) - *CC BY 4.0*

- 10.4: Rare Events, the Sample, Decision and Conclusion - *CC BY 4.0*
- 10.5: Additional Information and Full Hypothesis Test Examples - *CC BY 4.0*
- 11: Hypothesis Testing with Two Samples - *CC BY 4.0*
 - 11.1: Comparing Two Independent Population Means (Hypothesis test) - *CC BY 4.0*
 - 11.2: Comparing Two Independent Population Proportions (Hypothesis test) - *CC BY 4.0*
 - 11.3: Matched or Paired Samples - *CC BY 4.0*
- Back Matter - *CC BY 4.0*
 - Index - *CC BY 4.0*
 - Glossary - *CC BY 4.0*
 - Detailed Licensing - *Undeclared*
 - Index - *Undeclared*
 - Glossary - *Undeclared*
 - Detailed Licensing - *Undeclared*