

# INTRODUCTORY STATISTICS STA 2023



*Cheban Acharya*  
Florida Southwestern State College

# Introductor Statistics STA 2023

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](mailto:info@LibreTexts.org) or visit our main website at <https://LibreTexts.org>.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 08/24/2026

# TABLE OF CONTENTS

## Licensing

### 1: Introduction to Data

1.1: Data, Samples, and Populations

1.2: Displaying Data

### 2: Numerical Summary of Data- Center, Variation, and Location

2.1: Measures of Center of the Data

2.2: Measures of Spread of the Data

2.3: Measures of Location of the Data

### 3: Linear Regression and Correlation

3.1: Scatter Diagrams and Correlation

3.2: Least Squares Regression Line

### 4: Probability

4.1: Fundamental Concepts of Probability

4.2: The Addition and Complement Rules

4.3: Independence and Multiplication Rules

### 5: Discrete Probability Distributions

5.1: Discrete Probability Distributions

5.2: Binomial Probability Distributions

### 6: The Normal Distribution

6.1: Properties of the Normal Distribution

6.2: Normal Distribution and Applications

### 7: Sampling Distributions

7.1: The Sampling Distribution of the Sample Mean

7.2: The Sampling Distribution of Sample Proportion

### 8: Estimating the Value of a Parameter

8.1: Estimating Population Proportion

8.2: Estimating Population Mean

### 9: Hypothesis Testing

9.1: Language of Hypothesis Testing

9.2: Test for the Population Proportion

9.3: Test for the Population Mean

## Index

## Glossary

## Detailed Licensing



# Licensing

---

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

# 1: INTRODUCTION TO 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."

1.1: Data, Samples, and Populations

1.2: Displaying Data

## Contributors

Template:ContribOpenStax

---

This page titled [1: Introduction to 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.

# 1.1: Data, Samples, and Populations

---

## Introduction

Like most people, you probably feel that it is important to "take control of your life." But what does this mean? Partly, it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study.

Here are some claims you may have heard several times. (We are not saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentyne.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- Native Americans are significantly more likely to be hit crossing the streets than are people of other ethnicities.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study suggests that consuming egg whites may increase one's lifespan.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is a 70% chance that in a room full of 30 people, at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

Before going further, however, two distinct meanings of the word "statistics" should be clarified. First, "statistics" could mean the academic discipline, a field of study, involving the collection, presentation, interpretation, and use of data. When used in this sense, the word is considered singular (though it looks plural). Second, in a narrower sense, a number collected in a survey or study is referred to as a "statistic." For instance, the average height of a sample of 100 people would be a (one) statistic. This is a singular word (of course, two or more of these numbers would be referred to as "statistics"). Be aware of the difference when you encounter this word.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are; they come from psychology, health, law, sports, business, etc. Indeed, data and data-interpretation show up in discourse from virtually every facet of contemporary life.

Statistics are often presented to lend credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. (It is not, of course, the only step needed for this purpose.) This chapter will help you learn statistical essentials. It will make you into an intelligent consumer of statistical claims.

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to question the statistics that you encounter. The British Prime Minister Benjamin Disraeli famously said, "There are three kinds of lies -- lies, damned lies, and statistics." This quote reminds us why it is so important to understand statistics. So let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

We have put the emphasis on defending ourselves against fraudulent claims wrapped up as statistics. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. When a research team is testing a new treatment for a disease, statistics allow them to conclude based on a relatively small trial that there is good evidence their drug is effective. Statistics allowed prosecutors in the 1950's and 60's to demonstrate racial bias existed in jury panels. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases.

Before we begin gathering and analyzing data, we need to characterize the **population** we are studying. If we want to study the

amount of money spent on textbooks by a typical first-year college student, our population might be all first-year students at your college. Or it might be the following:

- All first-year community college students in the state of Washington.
- All first-year students at public colleges and universities in the state of Washington.
- All first-year students at all colleges and universities in the state of Washington.
- All first-year students at all colleges and universities in the entire United States.
- And so on.

## Components of Statistical Study

### Definition: Individual

In statistics, an individual is a single object or member of a population or sample that is being studied.

If a study records the heights of 30 students, **each student is an individual**, and **height** is a variable measured on those individuals.

### Definition: Population

In statistics, a population is the entire group of individuals or items that we want to study or draw conclusions about.

The population includes all individuals of interest, not just a subset.

Sometimes, the intended population is called the **target population**. If we design our study poorly, the collected data may not accurately represent the intended population.

Why is it important to specify the population? We might get different answers to our question depending on the population we are studying. First-year students at the University of Washington might take slightly more diverse courses than those at your college, and some of these courses may require less popular textbooks that cost more; or, on the other hand, the University Bookstore might have a larger pool of used textbooks, reducing the cost of these books to the students. Whichever the case (and it is likely that some combination of these and other factors is in play), the data we gather from your college will probably not be the same as that from the University of Washington. Particularly when conveying our results to others, we want to be clear about the population our data describe.

### Example 1.1.1: Describe Population

Identify the population

1. A newspaper website contains a poll asking people their opinion on a recent news article.
2. What is the proportion of drug users in prison populations in a certain country?
3. What percentage of people in Fort Myers favor casino gambling?

*Answer available at the end of this section.*

### Definition: Parameter

A **parameter** is a value (e.g., average, percentage) calculated from all the data in a population. In other words, a parameter is the numerical value that describes the population.

We seldom see parameters; however, since surveying an entire population is usually very time-consuming and expensive, unless the population is very small or we already have the data collected.

### Definition: Census and Survey

In a **census**, we gather information from every member of a population, and a **census** is a survey of an entire population. In a census, we gather information from every member of a population. In a **survey**, we gather information from a sample representing the population.

You are probably familiar with two common censuses: the official government Census that attempts to count the population of the U.S. every ten years, and voting, which asks the opinion of all eligible voters in a district. The first of these demonstrates one additional problem with a census: the difficulty in finding and getting participation from everyone in a large population, which can bias or skew the results.

There are occasional instances when a census is appropriate, typically when the population is relatively small. For example, if the manager of Starbucks wanted to know the average number of hours her employees worked last week, she should be able to pull up payroll records or ask each employee directly.

Since surveying an entire population is often impractical, we usually select a sample to study.

### Definition: Sample

A **sample** is a smaller subset of the entire population, ideally one that is fairly representative of the whole population.

We discuss sampling methods in greater detail in a later section. For now, let us assume that samples are chosen in an appropriate manner. If we survey a sample, say 100 first-year students at your college, and find the average amount of money spent by these students on textbooks, the resulting number is called a statistic.

### Definition: Statistic

A **statistic** is a value (average, percentage, etc.) calculated using the data from a sample. In other words, a statistic is the numerical value that describes the sample.

**Note:** Compare the statistic with the parameter.

### Your Turn 1.1.1: Identify the Components of Statistical Study

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 1.1.2: Parameter or Statistic

A researcher wanted to know how citizens of Tacoma felt about a voter initiative. To study this, she goes to the Tacoma Mall and randomly selects 500 shoppers and asks them their opinion. 60% indicates they are supportive of the initiative. What is the sample and population? Is the 60% value a parameter or a statistic?

*Answer available at the end of this section.*

### Definition: Variables

A variable in statistics is any characteristic, number, or quantity that can be measured, counted, or categorized and whose value varies from one entity to another (e.g., age, income, eye color, nationality, height).

### Your Turn 1.1.2: Identify the Components of Statistical Study

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Types of Data

Once we have gathered data, we might wish to classify it. Roughly speaking, data can be classified into two categories: qualitative data (also known as categorical data) and quantitative data.

### Definition: Qualitative and Quantitative Data

Categorical (qualitative) data are pieces of information that allow us to classify the objects under investigation into various categories. Qualitative data refers to non-numerical, descriptive information that captures qualities, characteristics, and meanings.

Quantitative data are numerical responses that allow meaningful arithmetic calculations. Quantitative data refers to information that can be measured, counted, and expressed numerically, offering objective statistical insights.

Quantitative data are of two types:

**Discrete data (Counted)** consists of distinct, separate values. You can think of these as "steps" on a ladder—you can be on one step or the next, but not in between. Discrete data are usually whole numbers; they cannot be broken into smaller fractions that make sense.

**Continuous data (Measured)** represents a scale that can be infinitely divided. It can take any value within a specific range. Continuous data can include decimals and fractions; it depends on the precision of the measuring instrument.

### Example 1.1.3: Identify the Data Type

1. We might conduct a survey to determine which movie each person in a math class has seen in a movie theater.
2. A survey could ask the number of movies they have seen in a movie theater in the past 12 months (0, 1, 2, 3, 4, ...)
3. Suppose we collect respondents' ZIP codes in a survey to track their geographic locations.

*Answer available at the end of this section.*

### Your Turn 1.1.3: Classify the Variables

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Types of Sampling Methods

As stated before, if you want to know something about a population, it is often impossible or impractical to examine the whole population. It might be too expensive in terms of time or money. It might be impractical – you can't test all batteries for their length of lifetime because there wouldn't be any batteries left to sell. You need to look at a sample. Hopefully, the sample behaves the same as the population.

When you choose a sample, you want it to be as similar to the population as possible. If you want to test a new painkiller for adults, you would want the sample to include people who are fat, skinny, old, young, healthy, not healthy, male, female, etc.

There are many ways to collect a sample. None are perfect, and you are not guaranteed to collect a representative sample. That is, unfortunately, the limitations of sampling. However, several techniques can result in samples that give you a semi-accurate picture of the population. Just remember to be aware that the sample may not be representative. As an example, you can take a random sample of a group of people that are equally males and females, yet by chance, everyone you choose is female. If this happens, it may be a good idea to collect a new sample if you have the time and money.

There are many sampling techniques, though only five will be presented here.

1. Simple Random Sample
2. Stratified Sample
3. Systematic Sample
4. Cluster Sample
5. Convenience Sample

### Definition: SRS

A **simple random sample (SRS)** of size  $n$  is a sample that is selected from a population in a way that ensures that every different possible sample of size  $n$  has the same chance of being selected. Also, every individual associated with the population has the same chance of being selected

1. Put all names in a hat and draw a certain number of names out.
2. Assign each individual a number and use a random number table or a calculator, or a computer to randomly select the individuals that will be measured.

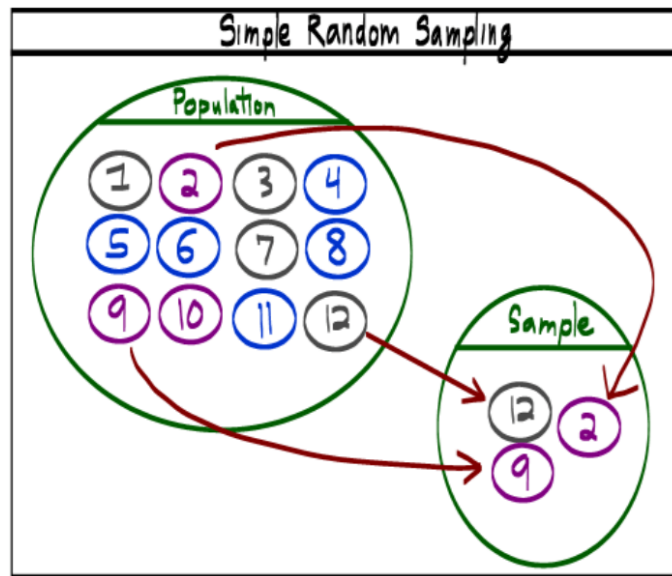


Figure 1.1.1: "Simple Random Sample Diagram" by Toros Berberyan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

In the simple random sampling, there is an original population of 12 balls. We conduct a random sample by randomly selecting three balls from the population to be part of the sample. In this case, we have picked the ones that have the numbers 2, 9, and 12.

### Definition: Stratified Sampling

**Stratified sampling** involves dividing the population into groups, called strata, and then selecting a simple random sample from each stratum.

1. If you want to look at musical preference, you could divide the individuals into age groups and then conduct simple random samples within each group.
2. To calculate the average price of textbooks, you could divide the individuals into groups by major and then conduct simple random samples within each group.

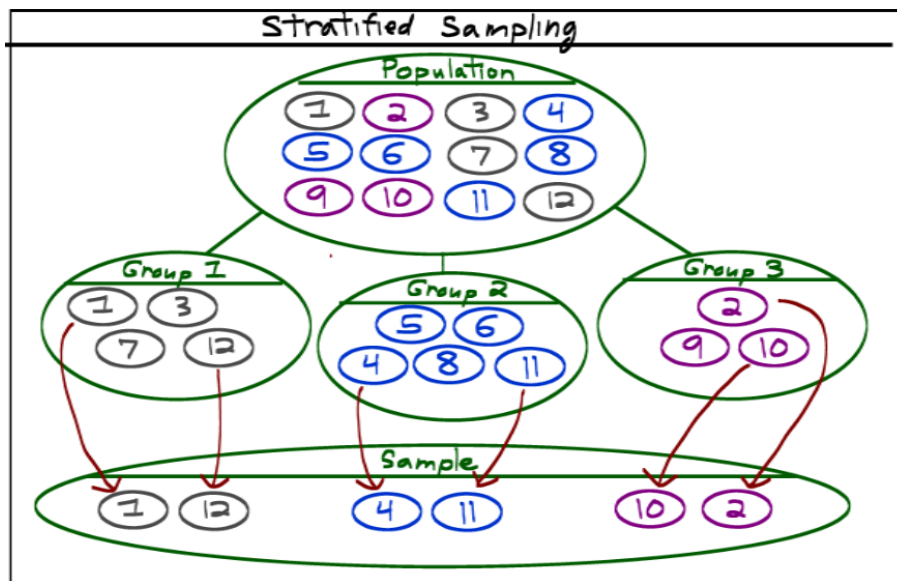


Figure 1.1.2: "Stratified Sampling Diagram" by Toros Berberyan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

The illustration above is an example of a stratified sample. There is an original population of 12 balls. They are grouped based on a common characteristic, in this case, color. We have 4 dark grey balls in the first group, 5 blue balls in the second group, and 3 purple balls in the third group. From **all groups, some members** are randomly selected to be part of the sample. From the first group, the balls that are numbered 1 and 12 are selected to be part of the sample. From the second group, the balls that are numbered 4 and 11 are selected to be part of the sample. From the third group, the balls that are numbered 2 and 10 are selected to be part of the sample. Based on stratified sampling, we ended up with a sample of 6 balls numbered 1, 2, 4, 10, 11, and 12.

### Definition: Systematic Sampling

**Systematic sampling** is where you randomly choose a starting place and then select every  $k$ th individual to measure.

1. You select every 5<sup>th</sup> item on an assembly line
2. You select every 10<sup>th</sup> name on the list
3. You select every 3<sup>rd</sup> customer that comes into the store.

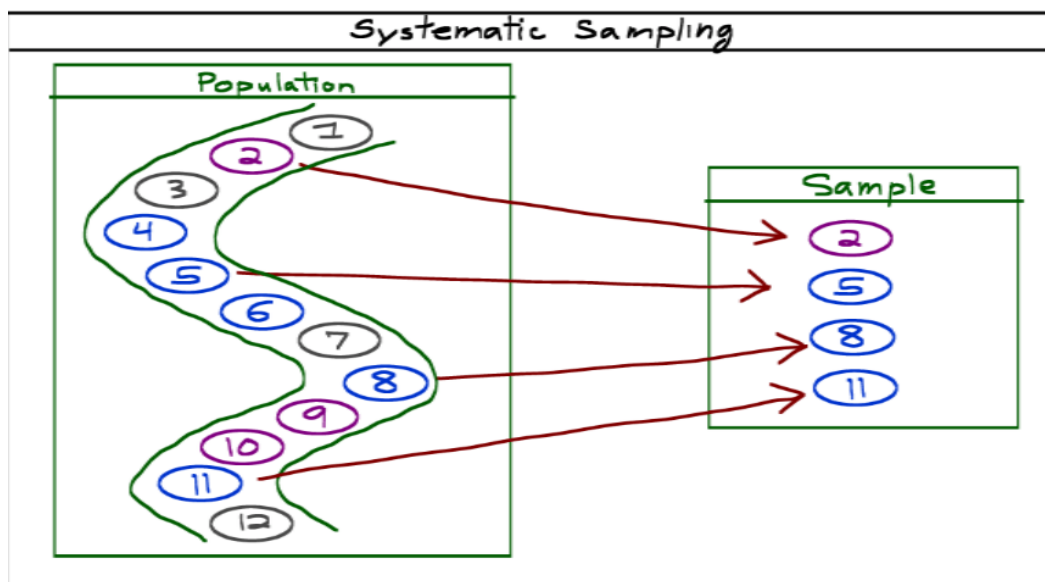


Figure 1.1.3: "Systematic Sampling Diagram" by [Toros Berberyan](#) (opens in new window) is licensed under [CC BY-SA 4.0](#) (opens in new window)

The diagram above represents a systematic sample. We have a population of 12 balls inside a curvy road. The first one is picked randomly. In this example, it will be the second one from above. Then, the rest of them are selected below the first one in a certain pattern. The diagram shows that every third one is selected to be part of the sample. Based on systematic sampling, the selected individuals in the sample are those numbered 2, 5, 8, and 11.

### Definition: Cluster Sampling

**Cluster sampling** involves dividing the population into groups, known as clusters. Randomly pick some clusters, then poll all individuals in those clusters.

1. A large city wants to poll all businesses in the city. They divide the city into sections (clusters), maybe a square block for each section, and use a random number generator to pick some of the clusters. Then, they poll all businesses in each chosen cluster.
2. You want to measure whether a tree in the forest is infected with bark beetles. Instead of having to walk all over the forest, you divide the forest into sectors and then randomly pick the sectors that you will travel to. Then record whether a tree is infected or not for every tree in that sector.

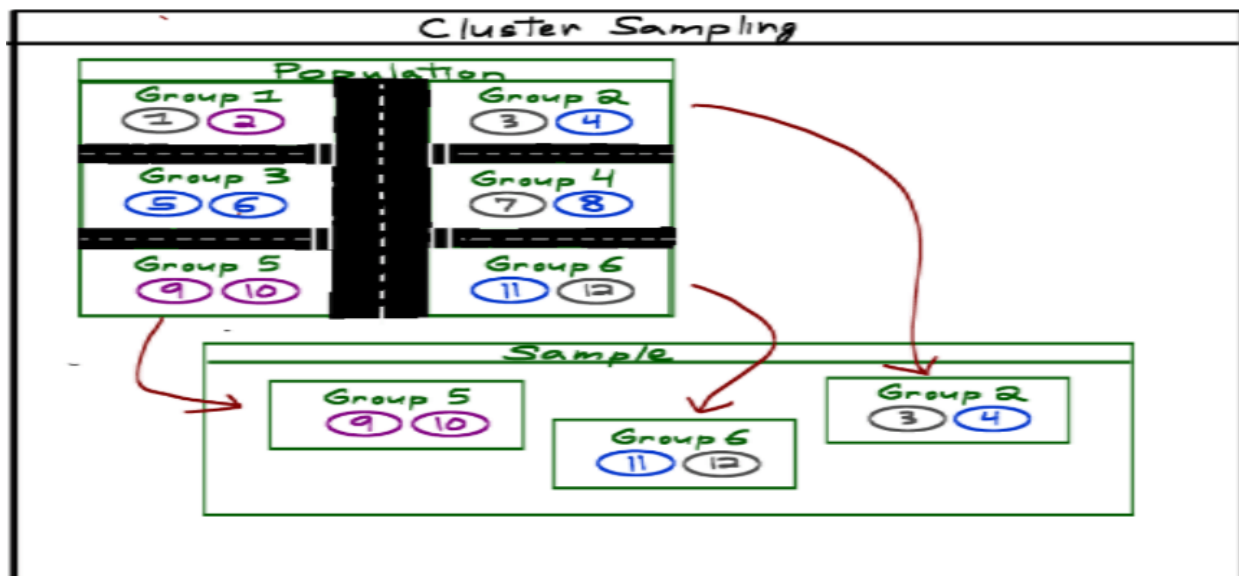


Figure 1.1.4: "Cluster Sampling Diagram" by [Toros Berberyan](#) (opens in new window) is licensed under [CC BY-SA 4.0](#) (opens in new window)

The illustration above is an example of a **cluster sample**. There is an original population of 12 balls. They are grouped into clusters (blocks). Each group has two balls. Then, **some groups** are selected, and **all members** from those groups are chosen to be part of the sample. In this example, the group that only had balls numbered 3 and 4 is selected to be part of the sample. Then, a second group with only balls numbered 9 and 10 is selected to be part of the sample. Finally, a third group with only balls numbered 11 and 12 is selected to also be part of the sample. Based on cluster sampling, we ended up with three groups and all its members being part of the sample. The sample included balls with the numbers 3, 4, 9, 10, 11, and 12.

#### Note: Cluster and Stratified?

Many people confuse stratified sampling and cluster sampling. In **stratified sampling**, you use **all the groups** and **some members** from each group. Cluster sampling is the opposite approach. **Cluster sampling** involves **selecting** a subset of **groups** and **all the members** within each group.

The four sampling techniques presented each have advantages and disadvantages. Another sampling technique is sometimes used because the researcher is unfamiliar with it or because it is easier to implement. This sampling technique is known as a convenience sample.

#### Definition: Convenience Sampling

**A convenience sample** is one where the researcher selects individuals to be included who are readily accessible to the researcher for collection.

Stopping shoppers in a mall to ask for feedback on a new product or posting a survey link on social media or a company newsletter are examples of convenience sampling.

If you want to know the opinion of people about the criminal justice system, you stand on a street corner near the county courthouse, and question the first 10 people who walk by. The people who walk by the county courthouse are most likely involved in some way with the criminal justice system, and their opinions would not necessarily represent the views of all individuals.

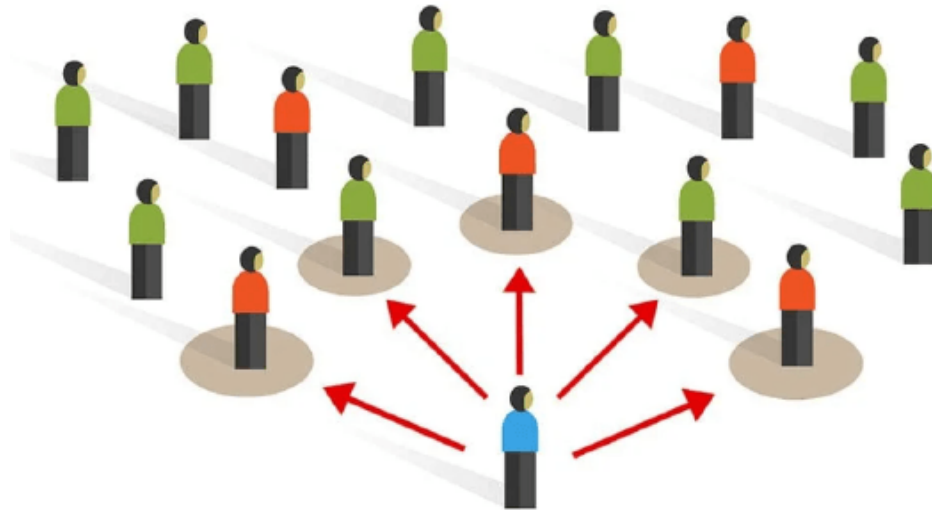


Figure 1.1.5: Convenience Sampling. (CS([opens in new window](#)))

#### Note: Is a Convenience Sample Representative?

A **convenience sample** is not representative and **should be avoided**. It has high sampling bias.

#### Example 1.1.4: Identify the Type of Sampling

Banner Health is a multi-state nonprofit chain of hospitals. Management wants to assess the incidence of complications after surgery. They wish to use a sample of surgery patients. Several sampling techniques are described below. Categorize each technique as a simple random sample, stratified sample, systematic sample, cluster sample, or convenience sampling.

1. Obtain a list of patients who underwent surgery at any Banner Health facility. Divide the patients into groups based on the type of surgery. Draw simple random samples from each group.
2. Obtain a list of patients who underwent surgery at any Banner Health facility. Number these patients, and then use a random number table to obtain the sample.
3. Randomly select some Banner Health facilities from each of the seven states, and then include all the patients on the surgery lists of the states.
4. At the beginning of the year, instruct each Banner Health facility to record any complications from every 100<sup>th</sup> surgery.
5. Instruct each Banner Health facility to record any complications from 20 surgeries this week and send in the results.

Answer available at the end of this section.

#### Your Turn 1.1.4: Identify the Type of Sampling Method

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Guidelines for Planning a Statistical Study

1. Identify the individuals that you are interested in studying. Realize that you can only make conclusions for these individuals. As an example, if you use a fertilizer on a certain genus of plant, you cannot say how the fertilizer will work on any other types of plants. However, if you diversify too much, then you may not be able to tell if there really is an improvement since you have too many factors to consider.
2. Specify the variable. You want to make sure the variable is something that you can measure, and make sure that you control for all other factors too. For example, if you are trying to determine if a fertilizer works by measuring the height of the plants on a particular day, you need to make sure you can control how much fertilizer you put on the plants (which is what we call a treatment), and make sure that all the plants receive the same amount of sunlight, water, and temperature.
3. Specify the population. This is important in order for you to know for whom and what conclusions you can make.
4. Specify the method for measuring or observing.
5. Determine if you are taking a census or a sample. If taking a sample, decide on the sampling method.
6. Collect the data.
7. Use appropriate descriptive statistics and make decisions using appropriate inferential statistics.
8. Note any concerns you have about your data collection methods and provide recommendations for the future.

### Your Turn 1.1.5: Basic Vocabulary

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Attribution:

3. Levine JD, Gordon NC, Smith R, Fields HL. (1981) Analgesic responses to morphine and placebo in individuals with postoperative pain. *Pain*. 10:379-89.

[LibreText Book](#)(opens in new window)

"1.2: Sampling Methods(opens in new window)" by Kathryn Kozak(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

[LibreText Samplign Method](#)(opens in new window)

[CS] [Simplypsychology](#)(opens in new window) [[www.simplypsychology.org](http://www.simplypsychology.org)]

## Answers

### Example 1.1.1: Describe Population

1. While the target (intended) population may have been all people, the real population of the survey is readers of the website.
2. All prisoners in that country.
3. All adult people in Fort Myers who are of legal gambling age.

### Example 1.1.2: Parameter or Statistic

The sample is the 500 shoppers questioned. The population is less clear. While the intended population of this survey was Tacoma citizens, the effective population was mall shoppers. There is no reason to assume that the 500 shoppers questioned would be representative of all Tacoma citizens.

The 60% value was based on the sample, so it is a statistic.

### Example 1.1.3: Identify the Data Type

1. When we conduct such a survey, the responses would look like: *Finding Nemo*, *The Avengers*, or *Dunkirk*. We might count the number of people who give each answer, but the answers themselves do not have any numerical values; we cannot perform computations with an answer like "*Finding Nemo*." This would be categorical data.
2. This would be quantitative data. Other examples of quantitative data would be the running time of the movie you saw most recently (104 minutes, 137 minutes,

104 minutes, ...) or the amount of money you paid for a movie ticket the last time you went to a movie theater (\$5.50, \$7.75, \$9, ...).

3. ZIP codes are numbers, but we can't do any meaningful mathematical calculations with them (it doesn't make sense to say that 98036 is "twice" 49018 — that's like saying that Lynnwood, WA is "twice" Battle Creek, MI, which doesn't make sense at all), so ZIP codes are really categorical data.

**Example 1.1.4: Identify the Type of Sampling**

1. This is a stratified sample because the patients were separated into strata, and random samples were then taken from each stratum. The problem is that some types of surgery may be more likely to cause complications than others. Of course, the stratified sample would show you this.
2. This is a random sample, since each patient has an equal chance of being selected. The problem with this one is that it will take a while to collect the data.
3. This is a cluster sample because all patients are questioned in the selected hospitals. The problem with this is that you might have, by chance, selected hospitals with no complications.
4. This is a systematic sample since they selected every 100<sup>th</sup> surgery. The problem with this is that if every 90<sup>th</sup> surgery has complications, you wouldn't see this come up in the data.
5. This is a convenience sample, since the facility left it up to them on how to do it. The problem with convenience samples is that the person collecting the data will probably collect data from surgeries that had no complications.

---

This page titled 1.1: Data, Samples, and Populations is shared under a license and was authored, remixed, and/or curated by Darlene Diaz (ASCCC Open Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.1: Populations and Samples by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH%20105/Math%20For%20Liberal%20Art%20Students%20(2017).pdf).
- 7: Equivalence Relations by Dave Witte Morris and Joy Morris is licensed CC BY-NC-SA 2.0.
- 8.3: Sampling Methods by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH%20105/Math%20For%20Liberal%20Art%20Students%20(2017).pdf).
- 6: Factoring by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.
- 8: Proof by Induction by Dave Witte Morris and Joy Morris is licensed CC BY-NC-SA 2.0.
- 8.5: Experiments by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH%20105/Math%20For%20Liberal%20Art%20Students%20(2017).pdf).

## 1.2: Displaying Data

### Displaying Qualitative Data

Categorical, or qualitative, data are pieces of information that allow us to classify the objects under investigation into various categories. We usually begin working with categorical data by summarizing it in a **frequency table**.

#### Definition: Frequency Table

A **frequency table** is a table with two columns. One column lists the categories, and another lists the frequencies with which the items in the categories occur (how many items fall into each category).

If the data we're visualizing is categorical, then we want a quick way to represent graphically the relative numbers of units that fall in each category. When we created the frequency distributions in the last section, all we did was count the number of units in each category and record that number (this was the *frequency* of that category). Frequencies are nice when we're organizing and summarizing data; they're easy to compute, and they're always whole numbers. But they can be difficult to understand for an outsider who's being introduced to your data.

#### Definition: Cumulative Frequency

**Cumulative frequency** is the **running total** of frequencies as you move through a data set from smallest to largest.

Cumulative Frequency is the sum of all frequencies up to that point.

Let's consider a quick example. Suppose you surveyed some people and asked for their favorite color. You communicated your results using a frequency distribution. Jerry is interested in data on favorite colors, so he reads your frequency distribution. The first row indicates that twelve people listed green as their favorite color. However, Jerry has no way of knowing if that's a lot of people without knowing how many people took your survey in total. Twelve is a pretty significant number if only twenty-five people took the survey, but it's next to nothing if you recorded a thousand responses. For that reason, we will often summarize categorical data not with frequencies, but with proportions. The proportion of data that falls into a particular category is computed by dividing the frequency for that category by the total number of units in the data.

#### Definition: Proportion and Relative Frequency

$$\text{Relative Frequency} = \frac{\text{Category frequency}}{\text{Total number of data}}$$

Proportions (Relative Frequency) can be expressed as fractions, decimals, or percentages. For example, relative frequency could be 0.35 or 35%.

#### Example 1.2.1: Finding Proportion

Recall the example in which a teacher recorded the responses on the first question of a multiple-choice quiz, with five possible responses (A, B, C, D, and E). The raw data were as follows.

*Response to Multiple Choice Questions*

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | A | C | A | B | B | A | E | A | C | A | A | A | C |
| E | A | B | A | A | C | A | B | E | E | A | A | C | C |

We computed a frequency distribution, relative frequency, and cumulative frequency.

Answer available at the end of this section.

### Your Turn 1.2.1: Complete Frequency and Relative Frequency Table

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Your Turn 1.2.2: Find Missing Values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



Sometimes we need an even more intuitive way of displaying data. This is where charts and graphs come in. There are many, many ways of displaying data graphically, but we will concentrate on one very useful type of graph called a bar graph. In this section, we will work with bar graphs that display categorical data; the next section will be devoted to graphs that display quantitative data.

## Bar Graph

### Definition: Bar Graph

A **bar graph** is a graph that displays a bar for each category, with the height of each bar indicating the frequency of that category.

Note: A bar graph is only used to display **categorical data**.

To construct a bar graph, we need to draw a vertical axis and a horizontal axis. The vertical direction will have a scale and measure the frequency of each category; the horizontal axis has no scale in this instance, but only the names of each category. The construction of a bar chart is most easily described by the use of an example.

An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get in accidents. To research this, they examine police reports for recent total-loss collisions. The data is summarized in the frequency table below.

Frequency and Relative Frequency Distribution Table

| Color | Frequency | Relative Frequency                  |
|-------|-----------|-------------------------------------|
| Blue  | 25        | $\frac{25}{216} = 0.1157 = 11.57\%$ |
| Green | 52        | $\frac{52}{216} = 0.2407 = 24.07\%$ |
| Red   | 41        | $\frac{41}{216} = 0.1898 = 18.98\%$ |

| Color | Frequency | Relative Frequency                  |
|-------|-----------|-------------------------------------|
| White | 36        | $\frac{36}{216} = 0.1667 = 16.67\%$ |
| Black | 39        | $\frac{39}{216} = 0.1806 = 18.06\%$ |
| Grey  | 23        | $\frac{23}{216} = 0.1065 = 10.65\%$ |
| Total | 216       | $1 = 100\%$                         |

Using our car data from above, note the highest frequency is 52, so our vertical axis needs to go from 0 to 52, but we might as well use 0 to 55, so that we can put a hash mark every 5 units:

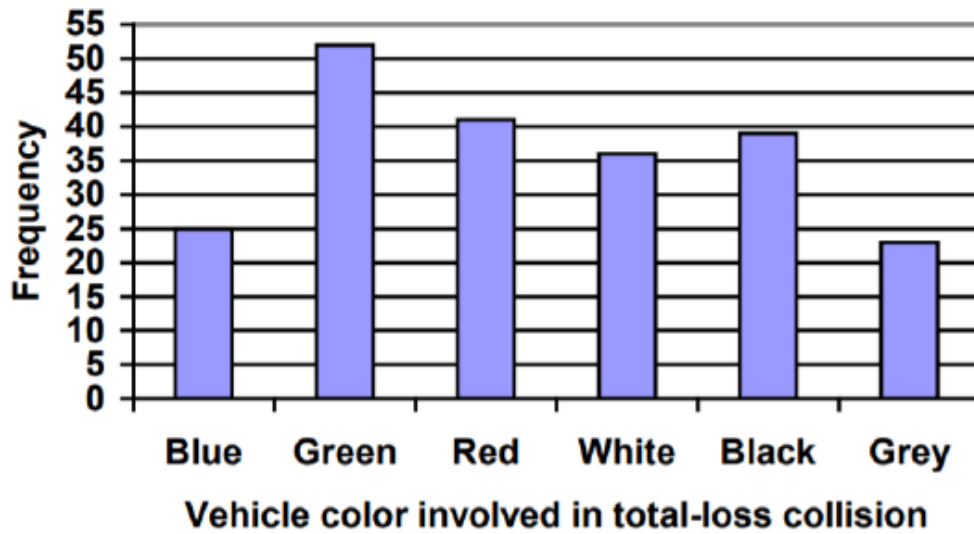


Figure 1.2.1: Bar Graph

Notice that the height of each bar is determined by the frequency of the corresponding color. The horizontal gridlines are a nice touch, but not necessary. In practice, you will find it useful to draw bar graphs using graph paper, where the gridlines are already in place, or by using technology. Instead of gridlines, we might also list the frequencies at the top of each bar, like this:

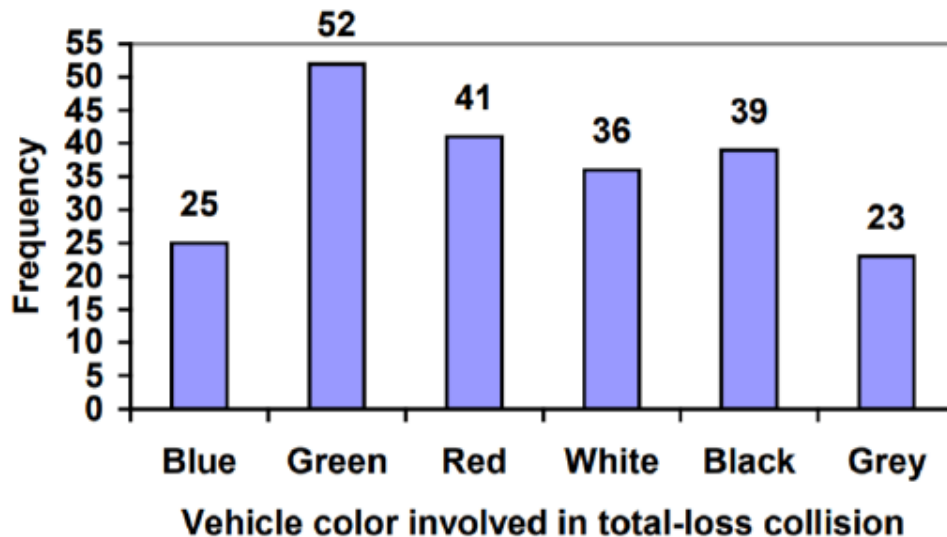


Figure 1.2.2: Bar Graph with Frequency Labelled

In this case, our chart might benefit from being reordered from largest to smallest frequency values. This arrangement can make

it easier to compare similar values in the chart, even without gridlines. When we arrange categories in decreasing frequency order, it is called a **Pareto chart**.

#### Definition: Pareto Chart

A **Pareto chart** is a bar graph ordered from highest to lowest frequency.

Transforming our bar graph from earlier into a Pareto chart, we get:

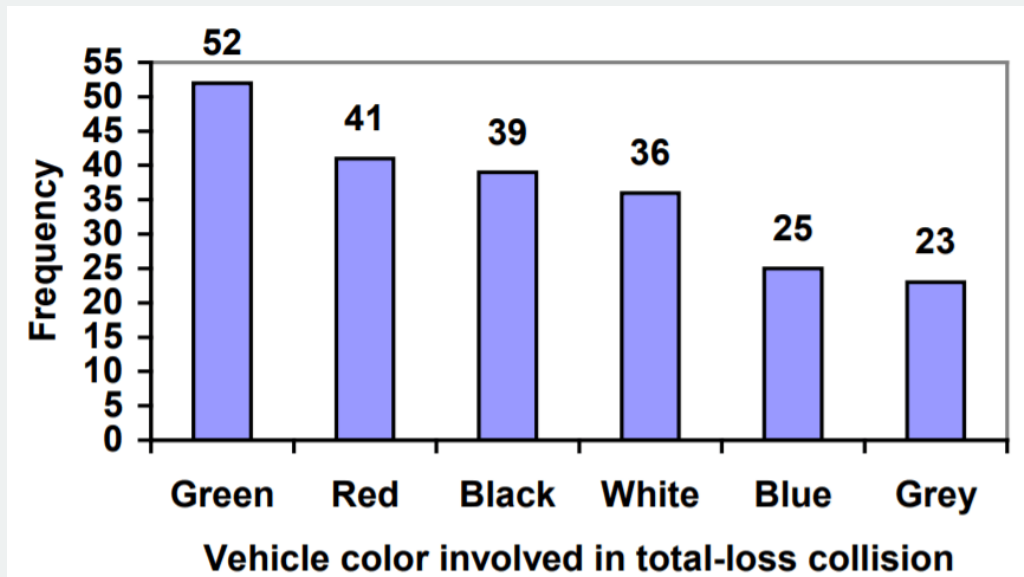


Figure 1.2.3: Pareto Chart

## Pie Chart, Pictograph, and Misleading Graph

Another way to represent qualitative data is with a **pie chart**. A pie chart is useful because it shows how a whole is divided into parts. It allows people to quickly compare proportions or percentages and easily see which categories make up the largest or smallest parts of the data. Pie charts are especially helpful when there are only a few categories and the goal is to show how each category contributes to the total (100%).

#### Definition: Pie Chart

A **pie chart** is a circle with wedges of varying sizes, arranged like slices of pie or pizza. The relative sizes of the wedges correspond to the relative frequencies of the categories.

#### Example 1.2.3: Pi Chart

For our vehicle color data in the above table, draw a pie chart.

*Answer available at the end of this section.*

Pie charts look nice but are harder to draw by hand than bar charts since to draw them accurately, we would need to compute the angle each wedge cuts out of the circle, then measure the angle with a protractor. Computers are much better suited to drawing pie charts. Common software programs like Microsoft Word or Excel, OpenOffice.org Write or Calc, or Google Docs are able to create bar graphs, pie charts, and other graph types. There are also numerous online tools that can create graphs [2].

### Your Turn 1.2.3: Create Bar Graph and Pie Chart

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



#### Definition: Pictograph

A **pictograph** (also known as a *pictogram*) is a type of chart that uses **images, icons, or symbols** to represent data.

A labor union might produce the graph to the right to show the difference between the average manager's salary and the average worker's salary.

Sometimes a pictograph is misleading.



Figure 1.2.6: Pictograph Manager Salaries Vs Worker Salaries

Looking at the figure 1.2.6, it would be reasonable to guess that the manager's salary is 4 times the worker's – the bag's area looks about \ (4 \) times as large. However, the manager's salaries are in fact only twice as large as the workers' salaries, as reflected in the picture by making the manager's bag twice as tall.

The following is another example of a pictograph.

# Books Read by Children

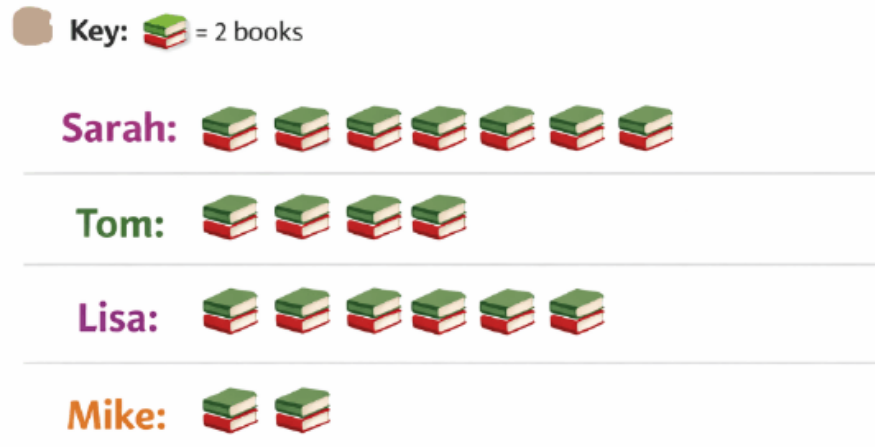


Figure 1.2.7: Pictograph: Books Read by Children

## Misleading Bar Graphs

Compare the two graphs below showing support for same-sex marriage rights from a poll taken in December 2008. The difference in the vertical scale on the first graph suggests a different story than the true differences in percentages; the second graph makes it look like three times as many people oppose marriage rights as support it.

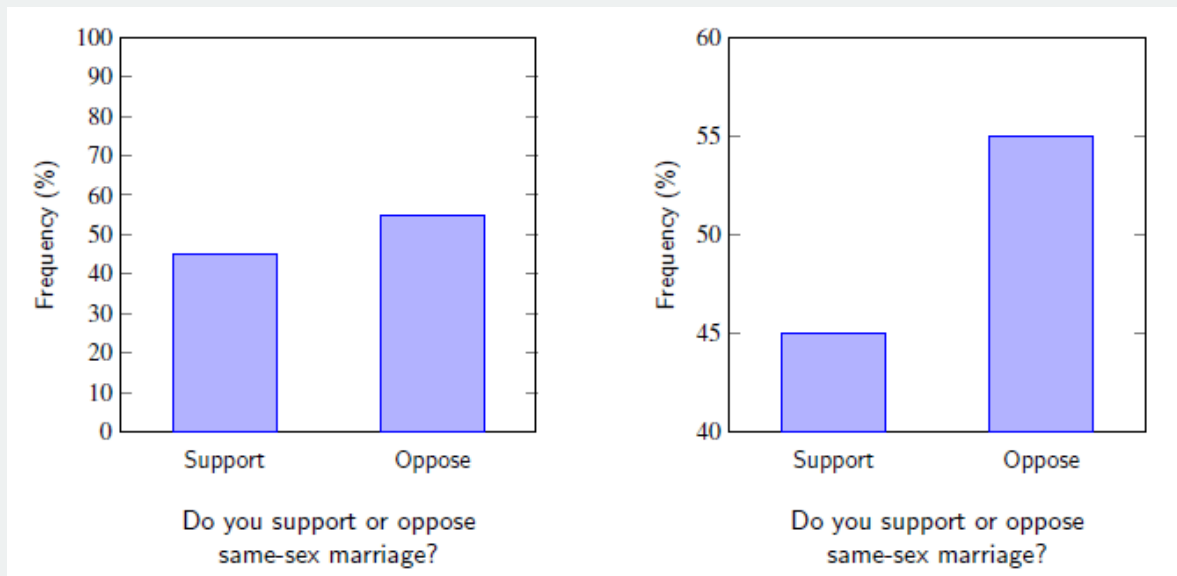


Figure 1.2.8: Misleading Bar Graph

## Misleading Pie Chart

A poll was taken asking people if they agreed with the positions of the 4 candidates for a county office. Does the pie chart accurately represent this data?

No: Because total percentage adds up to more than 100%

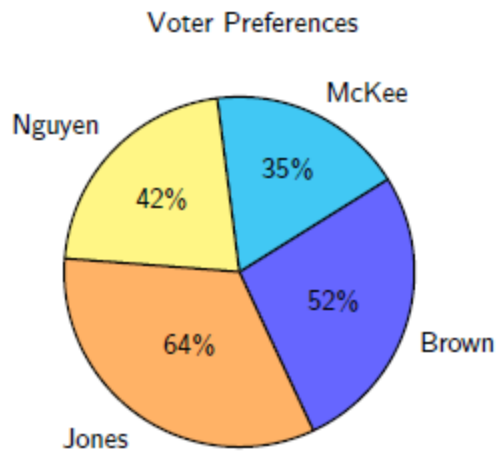


Figure 1.2.9: Misleading Pie Chart

## Displaying Quantitative Data

Quantitative, or numerical, data can also be summarized into frequency tables and graphs. A representation of quantitative data is called a **histogram**. The horizontal axis is a number

Some words we use to describe distributions are *uniform* (data are equally distributed across the range), *symmetric* (data are bunched up in the middle, then taper off in the same way above and below the middle), *left-skewed* (data are bunched up at the high end or larger values, and taper off toward the low end or smaller values), and *right-skewed* (data are bunched up at the low end, and taper off toward the high end). See the figures below.

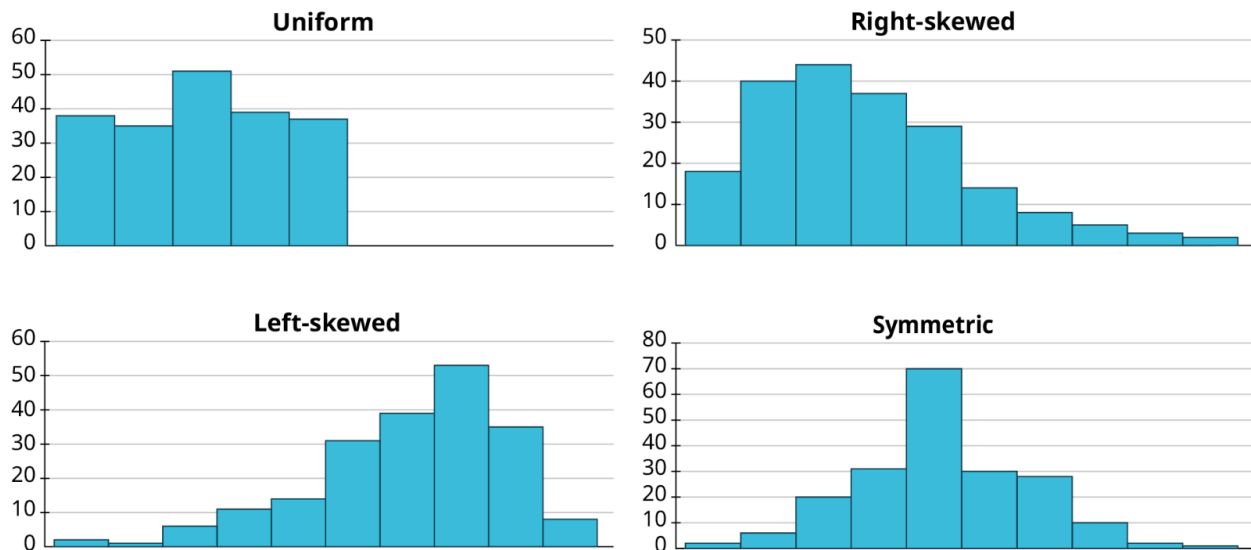


Figure 1.2.10: Distribution of Histogram

### Example 1.2.4: Construct Frequency and Relative Frequency Table

A teacher records scores on a 20-point quiz for the 30 students in his class. Note that the quiz score is given as whole numbers only. The scores are

19 20 18 18 17 17 19 17 20 18 20 16 20 15 17 12 18 19 18 19 17 20 18 16 15 18 20 5 0 0

These scores could be summarized into a frequency table by grouping like values:

Frequency and Relative Frequency Distribution Table

| Score | Frequency | Relative Frequency                |
|-------|-----------|-----------------------------------|
| 0     | 2         | $\frac{2}{30} = 0.067 = 6.7\%$    |
| 5     | 1         | $\frac{1}{30} = 0.033 = 3.3\%$    |
| 12    | 1         | $\frac{1}{30} = 0.033 = 3.3\%$    |
| 15    | 2         | $\frac{2}{30} = 0.067 = 6.7\%$    |
| 16    | 2         | $\frac{2}{30} = 0.067 = 6.7\%$    |
| 17    | 5         | $\frac{5}{30} = 0.1667 = 16.67\%$ |
| 18    | 7         | $\frac{7}{30} = 0.2333 = 23.33\%$ |
| 19    | 4         | $\frac{4}{30} = 0.1333 = 13.33\%$ |
| 20    | 6         | $\frac{6}{30} = 0.2000 = 20.00\%$ |
| Total | 30        | 1 = 100%                          |

Using this table, it would be possible to create a standard bar chart from this summary, as we did for categorical data:

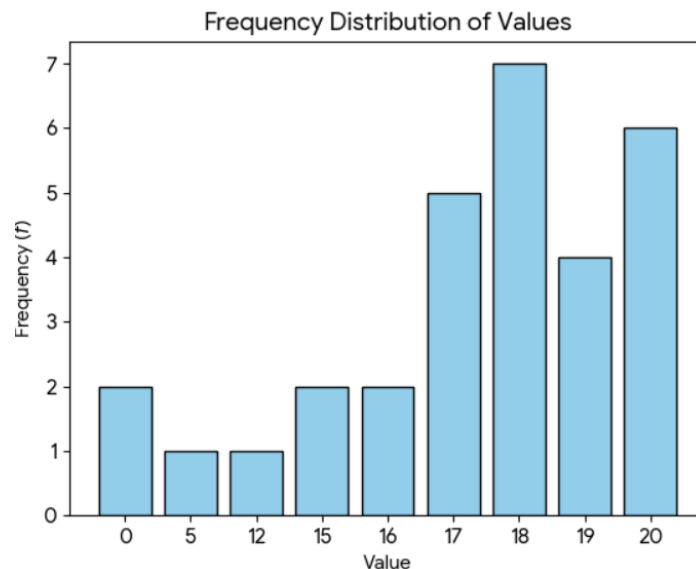


Figure 1.2.11: Bar Graph of 30 quiz scores

## Bar Graph Vs Histogram

### The Bar Graph (For Qualitative Data)

In a bar graph, each bar represents a separate category. For example, if you survey people on their favorite fruit, "Apple" and "Banana" are independent categories. There is no "in-between" value between an apple and a banana, which is why we leave spaces between the bars.

### The Histogram (For Quantitative Data)

A histogram groups numerical data into **bins** (class intervals). For example, if you are measuring the heights of people in a room, you might have a bin for "150 cm to 160 cm" and another for "160 cm to 170 cm." The bars touch because 160.10 cm follows immediately after 160 cm—the data flows continuously.

Comparison between bar graphs and histograms:

| Feature          | Bar Graph                      | Histogram                                                                               |
|------------------|--------------------------------|-----------------------------------------------------------------------------------------|
| <b>Data Type</b> | Categorical (qualitative) data | Numerical (quantitative) data                                                           |
| <b>Bars</b>      | Separated by gaps              | Usually bars touch each other (continuous data), but could be separated (discrete data) |

### Your Turn 1.2.4: Construct Frequency Distribution

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



If we have a large number of widely varying data values, creating a frequency table that lists every possible value as a category would lead to an exceptionally long frequency table and probably would not reveal any patterns. For this reason, it is common with quantitative data to group data into **class intervals**. A **class interval** is the range of values used to group data in a frequency distribution.

### Definition: Class Intervals

Class intervals are groupings of the data. In general, we define class intervals so that

- Each interval is equal in size. For example, if the first class contains values from 120 – 129, the second class should include values from 130 – 139.
- Each interval has a **lower** and an **upper limit**; e.g., for the interval 120 – 129, 120 is the lower limit and 129 is the upper limit.
- The **class width** is the difference between two consecutive lower limits.
- The **class width** is the same for every interval in the frequency table.
- We typically have between 5 and 20 classes, depending on the amount of data we're working with

Another way to represent data in Example 1.2.4 is in a group. The histogram of the data is also shown below.

| Class Interval<br>(Bin) | Frequency |
|-------------------------|-----------|
| 0–4                     | 2         |
| 4–8                     | 1         |
| 8–12                    | 0         |
| 12–16                   | 3         |
| 16–20                   | 18        |
| 20–24                   | 6         |

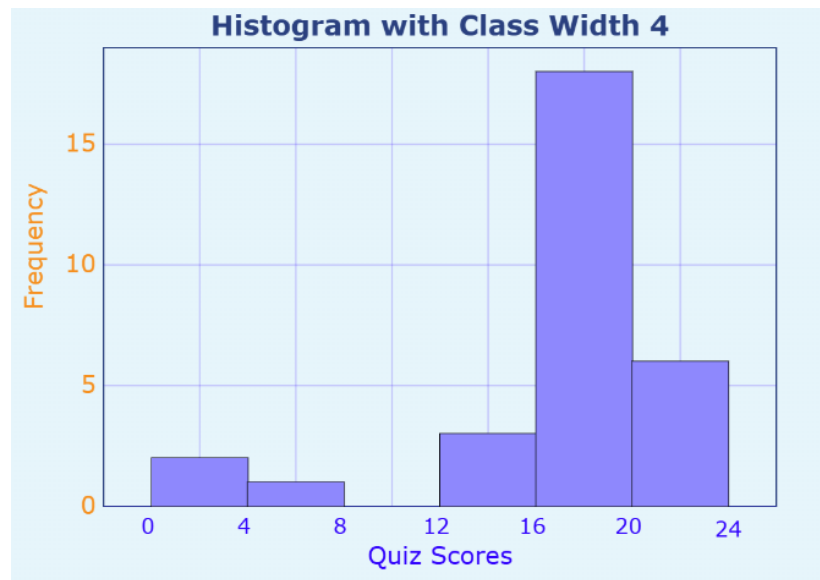


Figure 1.2.12: Histogram of 30 Quiz Scores

Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of  $263 - 121 = 142$ . We could create 7 intervals with a width of around 20, 14 intervals with a width of around 10, or somewhere in between. Often, we have to experiment with a few possibilities to find something that accurately represents the data. Let us try using an interval width of 15. We could start at 121, or at 120 since it is a nice round number.

Frequency Distribution Table for Data Organized in Classes

| Interval  | Frequency |
|-----------|-----------|
| 120 – 134 | 4         |
| 135 – 149 | 14        |
| 150 – 164 | 16        |
| 165 – 189 | 28        |
| 180 – 194 | 12        |
| 195 – 209 | 8         |
| 210 – 224 | 7         |
| 225 – 239 | 6         |
| 240 – 254 | 2         |
| 255 – 269 | 3         |

Notice, the class width is 15 since  $150 - 135 = 15$ ,  $165 - 150 = 15$ , and so on.

A histogram of this data would look like:

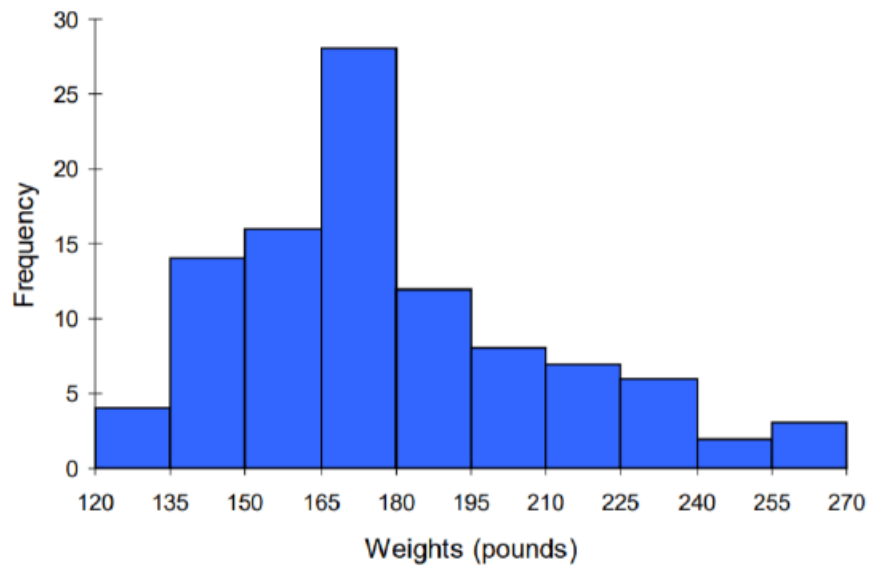


Figure 1.2.13: Histogram

### Your Turn 1.2.5: Meaning of Bin

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



In many software packages, you can create a graph similar to a histogram by putting the class intervals as the labels on a bar chart.

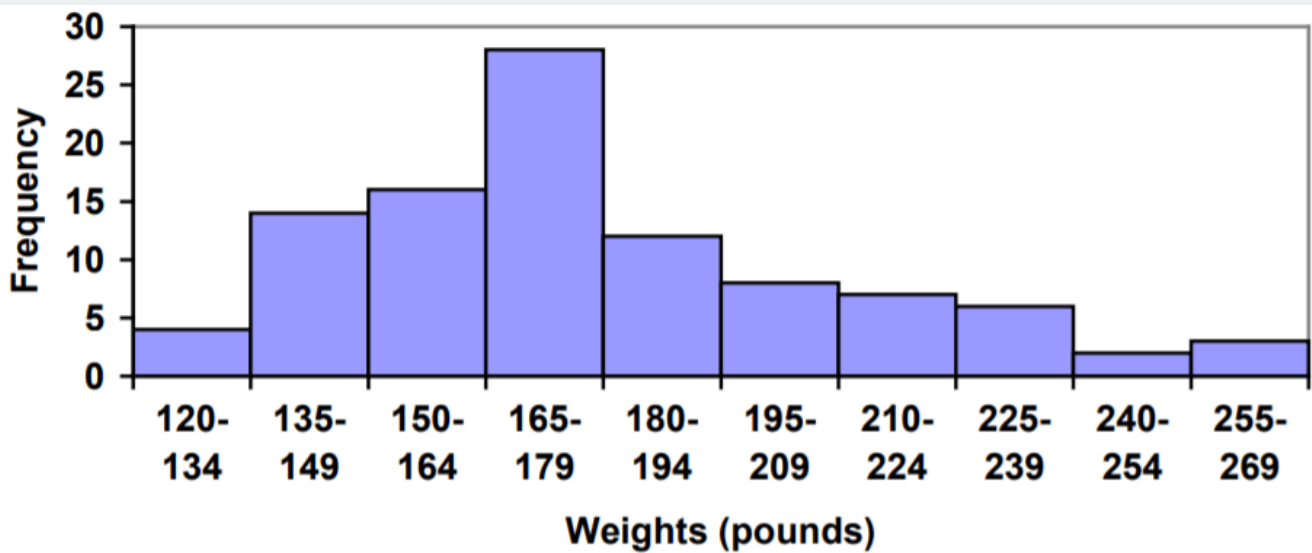


Figure 1.2.14: Histogram

Other graph types, such as pie charts, are possible for quantitative data. The usefulness of different graph types will vary depending on the number of intervals and the type of data being represented. For example, a pie chart of our weight data is difficult to read due to the large number of intervals we used.

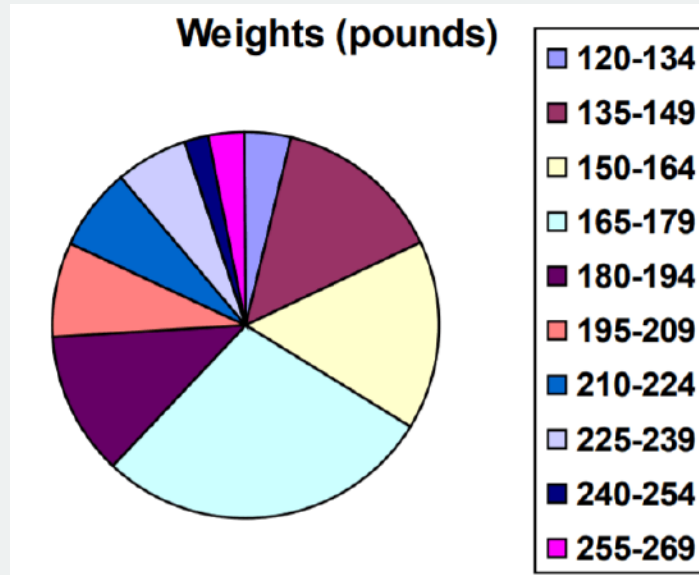


Figure 1.2.15: Pie Chart

### Your Turn 1.2.6: Complete Frequency and Relative Frequency Table

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Definition: Frequency Polygon

An alternative representation is a frequency polygon. A **frequency polygon** begins like a histogram, but instead of drawing a bar, a point is placed at the midpoint of each interval, with a height equal to the frequency.

Typically, the points are connected with straight lines to emphasize the distribution of the data. A **frequency polygon** is useful when we want to **compare two or more** groups on one chart.

This graph makes it easier to see that reaction times were generally shorter for the larger target and that the reaction times for the smaller target were more spread out.

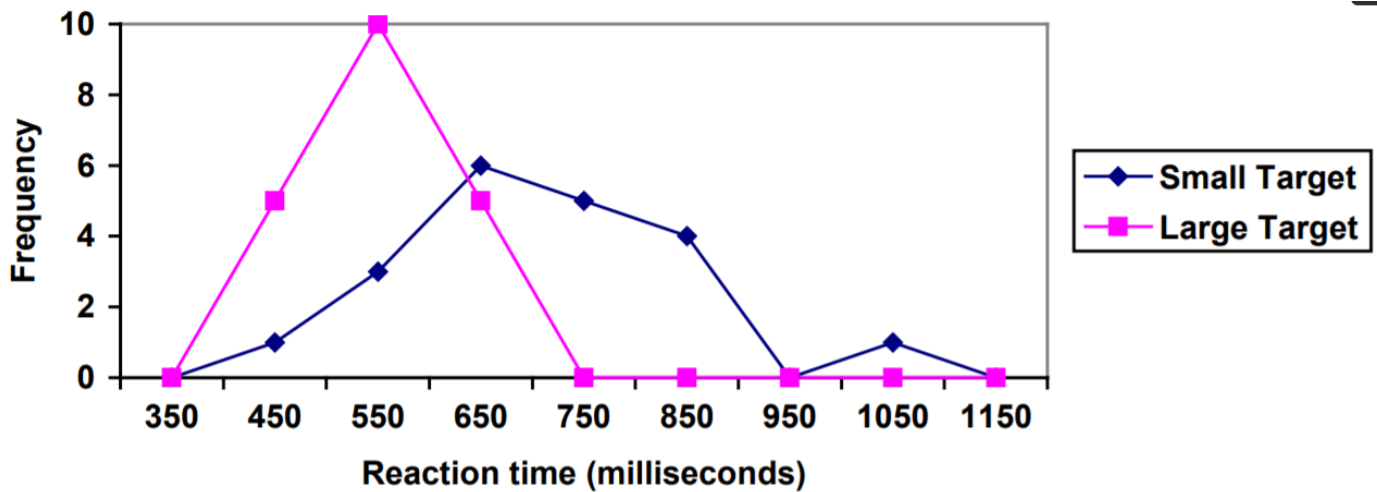


Figure 1.2.16: Frequency Polygon

### Your Turn 1.2.7: Read the Frequency Polygon

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Understanding Bar Graph

A bar chart is a visualization of categorical data that consists of a series of rectangles arranged side by side (but not touching). Each rectangle corresponds to a category. All of the rectangles have the same width. The height of each rectangle corresponds to either the number of units in the corresponding category or the proportion of the total units that fall into the category.

### Example 1.2.8: Understand Bar Graph

Students in a statistics class were asked to list their majors. A bar graph is drawn to visualize this data. Answer the following questions.

1. How many people participated in the survey?
2. How many more people plan to major in Biology than in Education?
3. Is the data on the horizontal axis categorical or quantitative?
4. Is this graph a bar chart, a histogram, or a Pareto chart?
5. What is the proportion of students who plan to major in Education?
6. What percentage of students are unsure about their major?

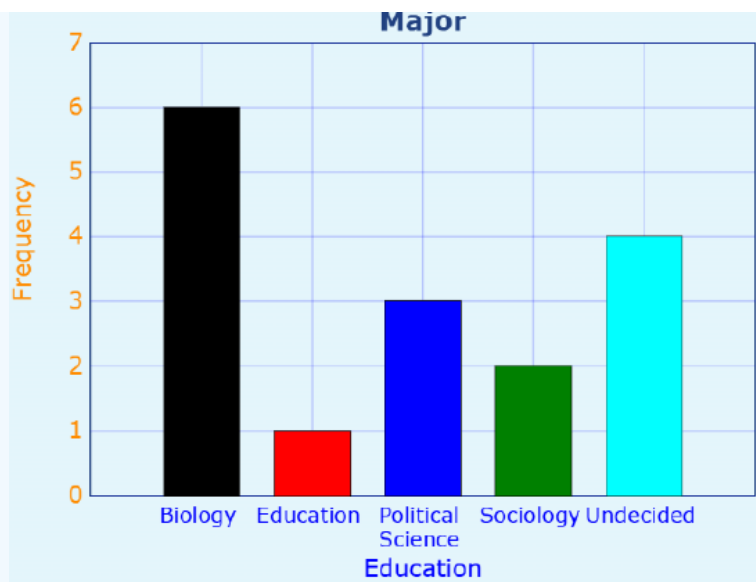


Figure 1.2.1: Bar Graph from Survey

Answer available at the end of this section.

### Your Turn 1.2.8: Average Monthly Precipitation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



If the data we're visualizing is categorical, we want a quick way to graphically represent the relative numbers of units in each category. When we created the frequency distributions in the last section, we simply counted the number of units in each category and recorded that count (the *frequency* of that category). Frequencies are nice when we're organizing and summarizing data; they're easy to compute, and they're always whole numbers. But they can be difficult to understand for an outsider who's being introduced to your data.

Let's get some practice reading bar graphs.

### Example 1.2.9: Reading Bar Graph

The bar graph shown gives data on 2020 model year cars available in the United States. Analyze the graph to answer the following questions.

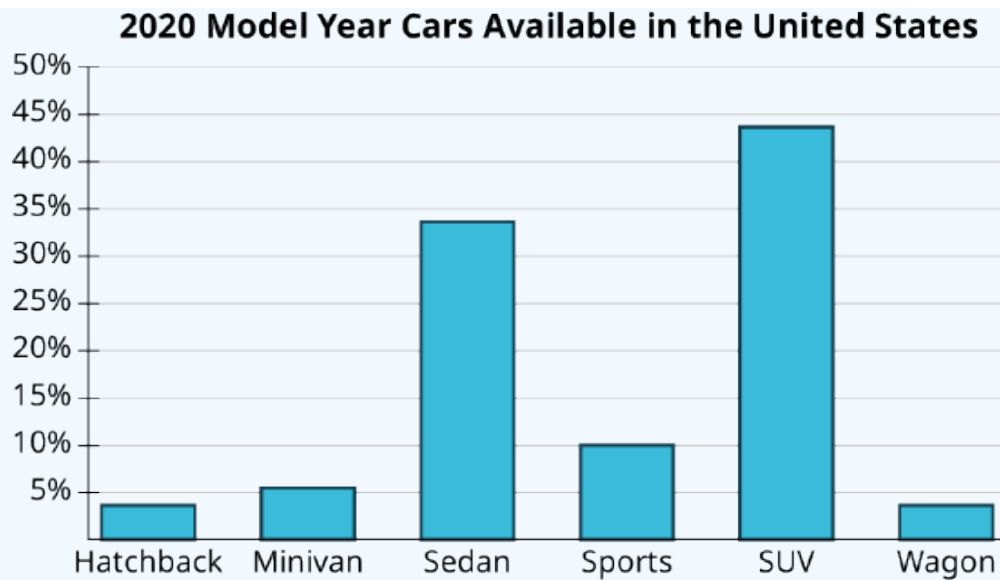


Figure 1.2.3 (data source: [consumerreports.org/cars](https://consumerreports.org/cars))

1. What proportion of available cars were sports cars?
2. What proportion of available cars were sedans?
3. Which categories of cars each made up less than 5% of the models available?

Answer available at the end of this section.

#### Your Turn 1.2.9: Reading Bar Graphs

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Understanding Pie Charts

A pie chart consists of a circle divided into wedges, with each wedge corresponding to a category. The proportion of the area of the entire circle that each wedge represents corresponds to the proportion of the data in that category. Pie charts are difficult to create without technology because they require careful angle measurements and precise circles, both of which are tasks better suited to computers.

#### Example 1.2.10: Construct Bar Graph

Use the data that follows to generate a pie chart.

Table 1.2.1 (data source: [www.consumerreports.org/cars](https://www.consumerreports.org/cars))

| Type  | Percent | Type      | Percent |
|-------|---------|-----------|---------|
| SUV   | 43.6%   | Minivan   | 5.5%    |
| Sedan | 33.6%   | Hatchback | 3.6%    |

| Type   | Percent | Type  | Percent |
|--------|---------|-------|---------|
| Sports | 10.0%   | Wagon | 3.6%    |

Answer available at the end of this section.

### Your Turn 1.2.10: Understanding Pie Chart

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### People in Mathematics: Florence Nightingale

Florence Nightingale (1820–1910) is best remembered today for her contributions in the medical field; after witnessing the horrors of field hospitals that tended to the wounded during the Crimean War, she championed reforms that encouraged sanitary conditions in hospitals. For those efforts, she is today considered the founder of modern nursing.

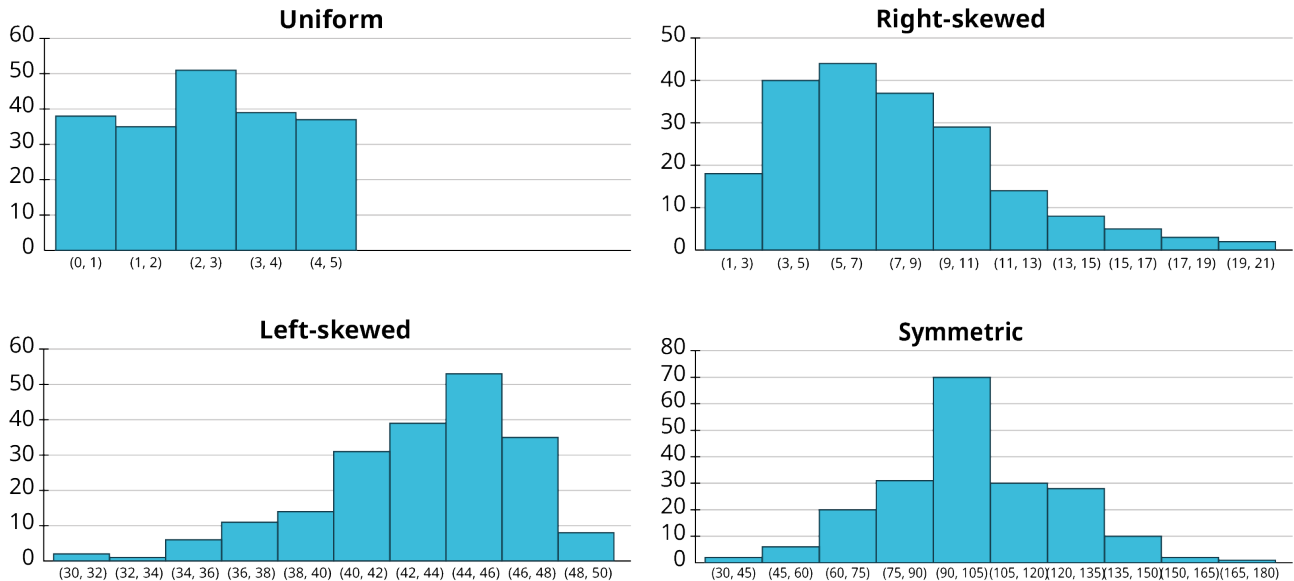


Figure 1.2.5: Florence Nightingale's significant contribution to the field of statistical graphics cannot be understated. (credit: "Florence Nightingale" by Library of Congress Prints and Photographs Division/Library of Congress (opens in new window) [hdl.loc.gov], public domain) Nightingale is also remembered for her contributions in statistics, especially in the ways we visualize data. She developed a version of the pie chart that is today known as a polar area diagram, which she used to visualize the causes of death among the soldiers in the war, highlighting the number of preventable deaths the British Army suffered in that conflict. In 1859, the Royal Statistical Society honored her for her contributions to the discipline by electing her to join the organization. She was the first woman to be so honored. She was later named an honorary member of the American Statistical Association. Nightingale's status as a revered pioneer in both nursing and statistics is a complex one, because some of her writings and opinions demonstrate a colonialist mindset and disregard for those who lost their lives and lands at the hands of the British. Her core statistical writings indicated that she felt superior to the Indigenous people she was treating. Members of both fields continue to debate her near-iconic role.

## More on Skewness of Histogram

As we see in the previous section, there are four distributions that describe quantitative data. They are: *uniform* (data are equally

distributed across the range), *symmetric* (data are bunched up in the middle, then taper off in the same way above and below the middle), *left-skewed* (data are bunched up at the high end or larger values, and taper off toward the low end or smaller values), and *right-skewed* (data are bunched up at the low end, and taper off toward the high end). See below figures.



1.2.6: Distribution of Histogram

Looking back at the stem-and-leaf plot in the previous example, we can see that the data are bunched up at the low end and taper off toward the high end; that set of data is right-skewed. Knowing the distribution of a set of data gives us useful information about the property that the data are measuring.

### Your Turn 1.2.11: Identify Distribution

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 1.2.12: Find Information from Histogram

The data in “AvgSAT”(opens in new window) contains the average SAT score for students attending every institution of higher learning in the US for which data is available. Create a histogram in Google Sheets of the average SAT scores. Use bins of width 50. Are the data uniformly distributed, symmetric, left-skewed, or right-skewed?

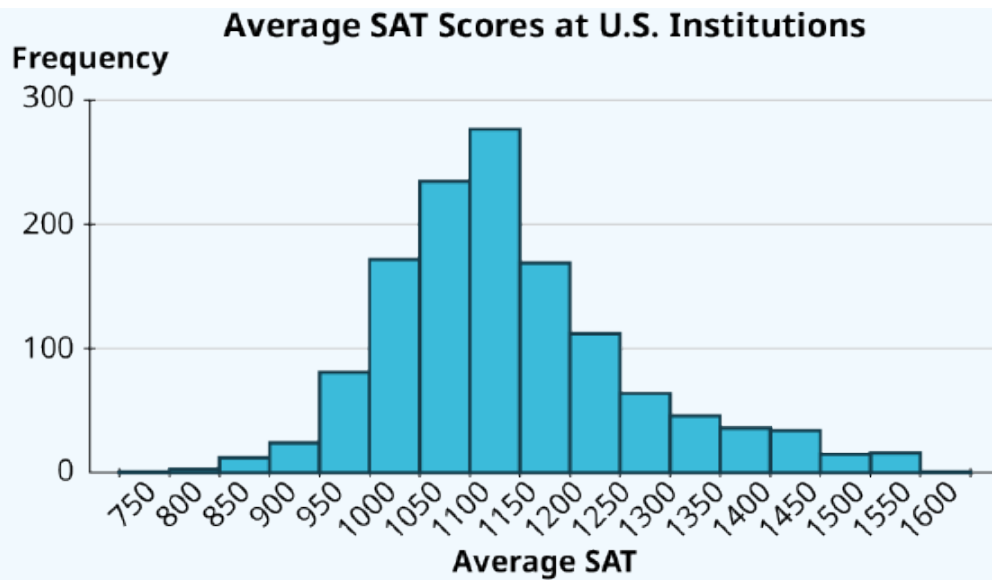


Figure 1.2.7: Fairly Symmetric Graph

Answer available at the end of this section.

#### Your Turn 1.2.12: Understanding Histogram

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Misleading Graphs

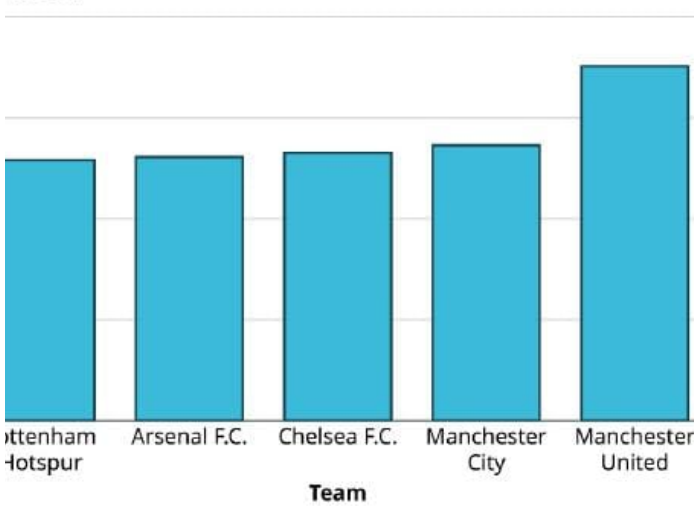
Graphical representations of data can be manipulated to intentionally mislead the reader. There are two primary ways this can be done: by adjusting the scales on the axes and by manipulating or misrepresenting the areas of the bars. Let's look at some examples of these.

#### Example 1.2.13: Misleading Bar Graph

The table below shows the teams and their payrolls in the English Premier League, the top soccer organization in the United Kingdom.

Top Five EPL Teams by Payroll (£1,000,000s), January 2020

Payroll (£1,000,000s)



Top Five EPL Teams by Payroll (£1,000,000s), Jan

Payroll (£1,000,000s)

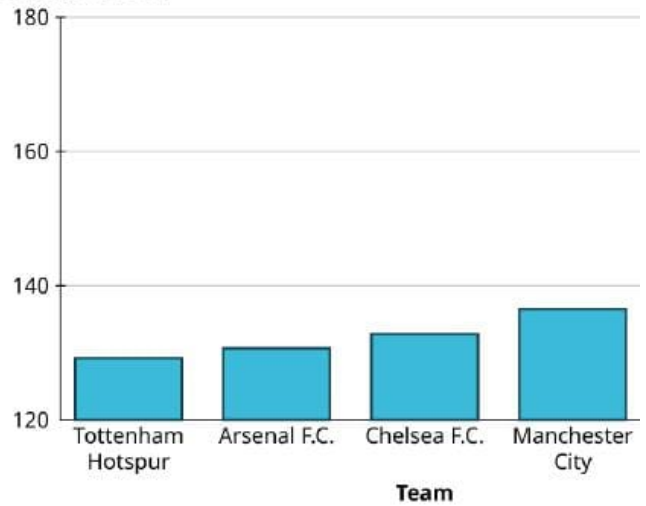


Figure 1.2.8: Misleading Bar Graph

How might someone present this data in a misleading way?

Answer available at the end of this section.

## Checkpoint

Beware of infographics! Areas overemphasize a difference that should be measured with a height!

Now, let's look at all 20 teams. This histogram indicates that the data are right-skewed, with the highest number of teams having a payroll between £40 million and £80 million:

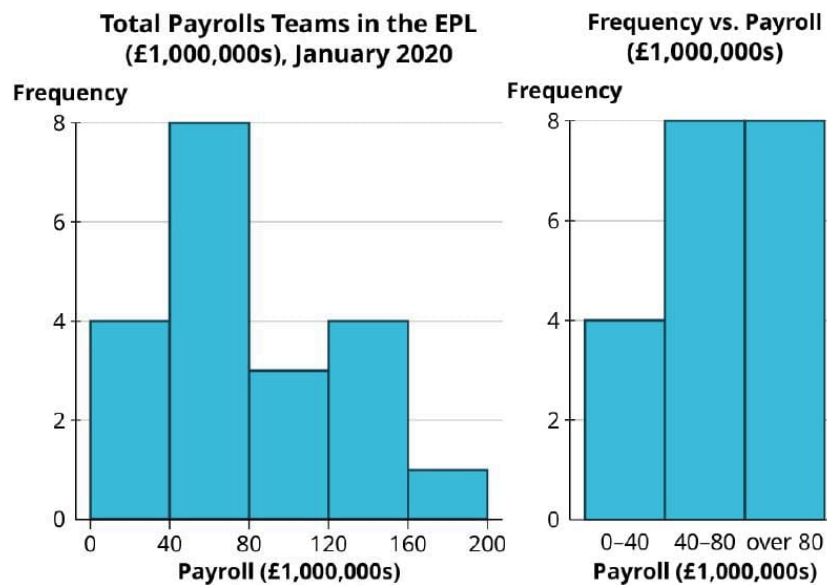


Figure 1.2.9: Misleading Bar Graphs

Even though this chart uses the same data, the skew seems to be reversed. Why? Well, even though this graph *looks* like a histogram, it isn't. Look closely at the labels on the horizontal axis; they don't correspond to spots on the axis, but instead provide a range, meaning this is a bar graph based on a binned frequency distribution.

When we review these ranges, we can see that the last range is misleading as it consists of all data “over 80.” If the bins all had the same width, that last bin would run from 80 to 120. However, the histogram shows that the maximum value for this data is between 160 and 200. If the last bin in this bar graph were labeled honestly, it would read “80–200,” which would drive home the fact that the bar’s width is misleading.

### Your Turn 1.2.13: Misleading Graph

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



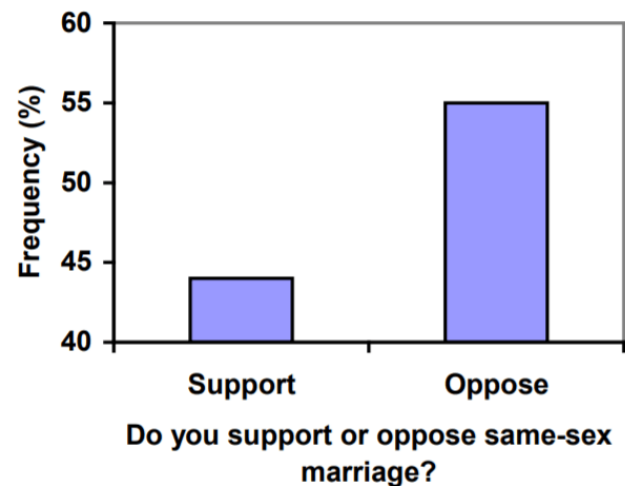
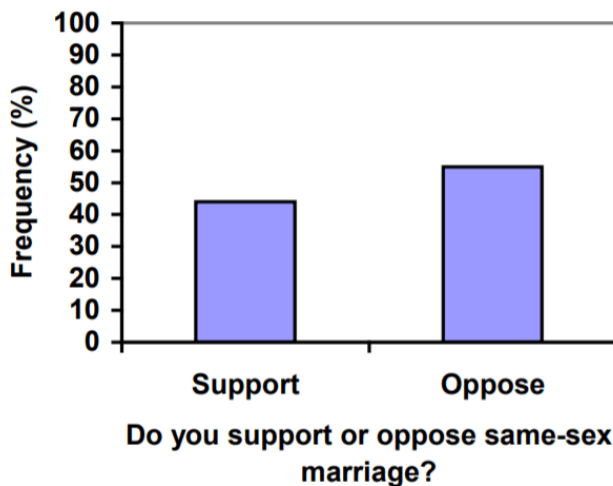
### Checkpoint

Always check the horizontal axis on histograms! The widths of all the bars should be equal.

Minard’s chart is remarkable in that it shows not only how the size of Napoleon’s army shrank drastically over time, but also its location on the map, the direction it was traveling at the time, and the temperature during the retreat.

### Example 1.2.14: Misleading Bar Graph

Compare the two graphs below showing support for same-sex marriage rights from a poll taken in December 2008 [3]. The difference in the vertical scale on the first graph suggests a different story than the true differences in percentages; the second graph makes it appear as though twice as many people oppose marriage rights as support them.



## Your Turn 1.2.14: Misleading Pie Chart

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Notes:

[Gallup Poll \[www.pollingreport.com\]](http://www.pollingreport.com)

[Creat a Graph NCES \[nces.ed.gov\]](http://nces.ed.gov)

[CNN poll \[www.pollingreport.com\]](http://www.pollingreport.com)

## Answers

### Example 1.2.1: Finding Proportion

To compute a proportion, we need the frequency (which is provided in the table above) and the total number of units represented in our data. We can find that by adding up the frequencies from all the categories:  $0 + 14 + 4 + 4 + 6 = 28$ .

To find the proportions, we divide the frequency by the total. For the first category ("A"), the proportion is  $\frac{14}{28} = 0.50 = 50\%$ . We can compute the other proportions similarly, filling in the rest of the table:

*Frequency and Relative Frequency Distribution Table*

| Category of Data | Frequency | Proportion(Relative Frequency)  | Cumulative Frequency |
|------------------|-----------|---------------------------------|----------------------|
| D                | 0         | $\frac{0}{28} = 0 = 0\%$        | 0                    |
| A                | 14        | $\frac{14}{28} = 0.50 = 50\%$   | $0 + 14 = 14$        |
| B                | 4         | $\frac{4}{28} = 0.143 = 14.3\%$ | $14 + 4 = 18$        |
| E                | 4         | $\frac{4}{28} = 0.143 = 14.3\%$ | $18 + 4 = 22$        |
| C                | 6         | $\frac{6}{28} = 0.214 = 21.4\%$ | $22 + 6 = 28$        |

Check your work: If you add up your proportions, you should get 1 or 100%. And frequency should be added to the total number of data in the data set.

### Example 1.2.3: Pi Chart

We use relative frequency to draw a pie chart.

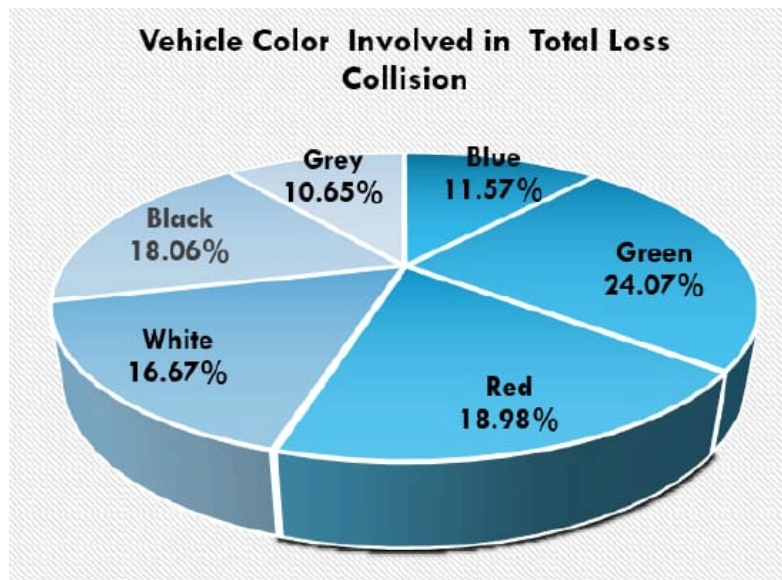


Figure 1.2.4: Pi Chart

#### Example 1.2.8: Understand Bar Graph

1.  $6 + 1 + 3 + 2 + 4 = 16$
2.  $6 - 1 = 5$  More students majoring in Biology than in Education.
3. Categorical (Qualitative)
4. Not Histogram. Not a Pareto chart.
5.  $\frac{1}{16} = 0.0625$
6.  $\frac{4}{16} = 0.25 = 25\%$

#### Example 1.2.9: Reading Bar Graph

1. The bar for sports cars goes up to 10%, so the proportion of models that are considered sports cars is 10%.
2. The bar corresponding to sedan goes up past 30% but not quite to 35%. It looks like the proportion we want is between 33% and 34%.
3. We're looking for the bars that don't make it all the way to the 5% line. Those categories are hatchback and wagon.

#### Example 1.2.10: Construct Bar Graph

First, enter the chart above into a new sheet in Google Sheets. Next, click and drag to select the full table (including the header row). Click on the "Insert" menu, then select "Chart." The result may be a pie chart by default; if it isn't, you can change it to a pie chart using the "Chart type" drop-down menu in the Chart Editor.

### 2020 Model Year Cars Available in the United States

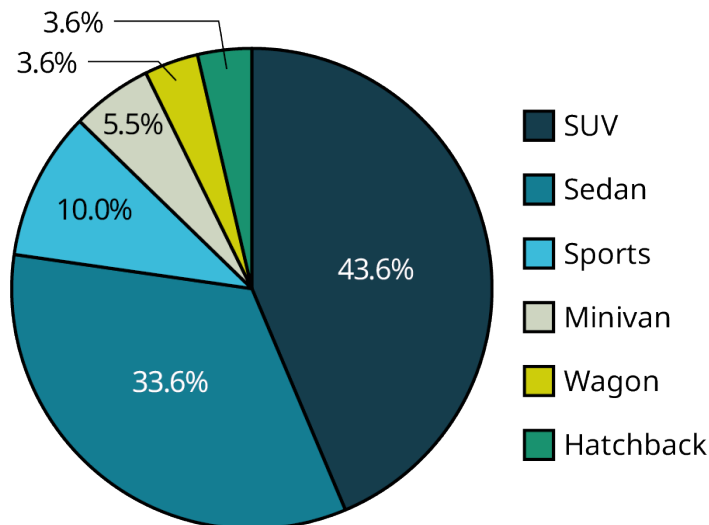


Figure 1.2.4 (data source: [consumerreports.org/cars](https://consumerreports.org/cars)) You can choose to use a legend to identify the categories, as well as label the slices with the relevant percentages.

#### Example 1.2.12: Find Information from Histogram

The data are fairly symmetric, but slightly right-skewed.

#### Example 1.2.13: Misleading Bar Graph

You should notice that despite using the same data, these two graphs look strikingly different. In the second graph, the gap between Manchester United and the other four teams appears significantly larger than in the first. The scale on the vertical axis has been manipulated here. The first graph's axis starts at zero, while the lowest value on the second graph's axis is 120. This trick has a strong impact on the viewer's perception of the data.

---

This page titled [1.2: Displaying Data](#) is shared under a license and was authored, remixed, and/or curated by LibreTexts.

- [9.1: Presenting Categorical Data Graphically](#) by Darlene Díaz is licensed CC BY-NC-SA 4.0. Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH%20105/Math%20For%20Liberal%20Art%20Students%20(2017).pdf).
- [8.2: Visualizing Data](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/contemporary-mathematics>.
- [9.2: Presenting Quantitative Data Graphically](#) by Darlene Díaz is licensed CC BY-NC-SA 4.0. Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH%20105/Math%20For%20Liberal%20Art%20Students%20(2017).pdf).

## 2: NUMERICAL SUMMARY OF DATA- CENTER, VARIATION, AND LOCATION

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.

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

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

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

### Contributors

- [Template:ContribOpenStax](#)

---

This page titled [2: Numerical Summary of Data- Center, Variation, and Location](#) 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: Measures of Center of the Data

### Mean, Median, and Mode

The **measure of center** (also known as central tendency) is a single value that attempts to describe a set of data by identifying the central position within that dataset. In statistics, the three most common measures are the **mean**, **median**, and **mode**.

Each measure provides a different perspective on what "typical" looks like, and the best one to use often depends on the shape of your data.

Let's begin by trying to find the most "typical" value of a data set.

Note that we just used the word "typical" although in many cases you might think of using the word "average." We need to be cautious with the word "average" as it has different meanings to different people in various contexts. One of the most common uses of the word "average" is what mathematicians and statisticians call the **arithmetic mean**, or just plain old **mean** for short. "Arithmetic mean" sounds rather fancy, but it is likely you have calculated a mean many times without realizing it; the mean is what most people think of when they use the word "average".

#### Definition: Mean

The **mean** of a set of data is the sum of the data values divided by the number of values.

#### Notation: Sample or Population Mean

The following are the symbols to represent the sample and the population mean

Sample mean:  $\bar{x}$  (Read as "x-bar"):  $\bar{x}$  is statistic.

Population mean:  $\mu$  (Read as "mew"):  $\mu$  is parameter.

#### Example 2.1.1: Find Mean

Marci's exam scores for her last math class were: 79, 86, 82, 94. The mean of these values would be:

$$\frac{79 + 86 + 82 + 94}{4} = 85.25$$

Typically, we round the mean to one more decimal place than the original data. In this case, we would round 85.25 to 85.3. Thus, we can say Marci's average score on her math exams was 85.25 or about 85.3.

Let's look at the following two examples in the box, and how the mean changes with data value(s)

The 15 families in a particular neighborhood are asked about their annual household income, rounded to the nearest \$1,000.

25 15 35 15 30 15 20 20 50 50 45 50 40 50 50

Calculating the mean annual household income

$$\bar{x} = \frac{25 + 15 + 35 + 15 + 30 + 15 + 20 + 20 + 50 + 50 + 45 + 50 + 40 + 50 + 50}{15} = 34$$

The mean household income of our sample is 34 thousand dollars (\$34,000).

Building on the last example, suppose a new family moves into the neighborhood, with a household income of \$2 million (\$2,000,000). Adding this to our sample, our mean (after rounding to the nearest dollar) is now:

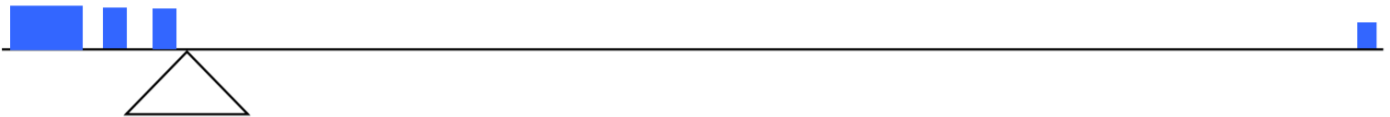
$$\bar{x} = \frac{25 + 15 + 35 + 15 + 30 + 15 + 20 + 20 + 50 + 50 + 45 + 50 + 40 + 50 + 50 + 2,000}{16} = 157$$

The mean household income of our sample is 157 thousand dollars (\$157,000).

Imagine the data values on a see-saw or balance scale. The mean is the value that balances the data, as shown in the picture below.



If we graph our household data, the \$2 million data value is so far out to the right that the mean has to adjust up to keep things in balance. The \$2 million data value is called an outlier.



For this reason, when working with data that has **outlier** values far outside the primary grouping, it is common to use a different measure of center, the **median**.

### What are outliers?

An outlier is a data value that is very different from the rest of the data and is far enough from the center.

### Definition: Median

The **median** of a set of data is the middle value when the data are in order. To find the **median**, follow these simple steps:

1. Step 1: Arrange the data. Put all the data values in order from smallest to largest.
2. Step 2: Count the number of values.
3. Step 3: Find the middle.

Let  $n$  be the total number of data values.

If  $n$  is odd:

The median is the middle number of the ordered data set (there is only one middle value). It is the data value in the  $(\frac{n+1}{2})^{\text{th}}$  position or ordered data set.

If  $n$  is even:

The median is the average of the two middle numbers. These are the values in the  $\frac{n}{2}^{\text{th}}$  and  $(\frac{n}{2} + 1)^{\text{th}}$  positions of the ordered data set.

We can interpret the median as “half of the data is less than the median, and the other half is more than the median.” Of course, we can rewrite this in the context of the problem.

Look at the following ordered 31 data.

37 33 33 32 29 28 28 23 22 22 22 21 21 21 20  
20 19 19 18 18 18 18 16 15 14 14 14 12 12 9 6

Since  $n = 31$  data values, an odd number, there is only one middle number, which is  $(\frac{31+1}{2} = 16)^{\text{th}}$  data value, which is 20.

So there are 15 values below 20 and 15 above it.

### Example 2.1.2: Find the Median

Find the median of these quiz scores:

5 10 8 6 4 8 2 5 7 7

*Answer available at the end of this section.*

Let's think about the previous example. When we added the 16<sup>st</sup> family's income, the mean was \$157,000 from \$34,00. That's a big difference in the average household income. We observe that the mean is influenced by the values of the data; that is, the mean can increase or decrease depending on the values of the data. However, when calculating the median including the 16<sup>st</sup> family's income, the median wasn't influenced at all.

In fact, the median is generally considered a better statistic for household income, as there is a wide range of income among families. Thus, the values of the data influence the mean, but not the median.

In addition to the mean and the median, there is one other common measurement of the “typical” value of a data set: the **mode**.

### Definition: Mode

The **mode** is the observed value of the data set that occurs most frequently.

The mode is most commonly used for categorical data, for which the median and mean cannot be computed. Additionally, the mode is the only central tendency used for both categorical and quantitative data. The mean and median are only used with quantitative data.

### Example 2.1.3: Finding Mode

1. The 15 families in a particular neighborhood are asked about their annual household income, rounded to the nearest \$1,000. Find the mode.

25 15 35 15 30 15 20 20 50 50 45 50 40 50 50

2. Five real estate exam scores are given below. Find the mode.

430, 430, 480, 480, 495

*Answer available at the end of this section.*

A data set can have more than one mode if several categories have the same frequency, or no modes if every category occurs

only once.

#### Your Turn 2.1.4: Find Mean, Median, and Mode

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Which Is Better: Mode, Median, or Mean?

If the mode, median, and mean all purport to measure the same thing (centrality), why do we need all three? The answer is complicated, as each measure has its own strengths and weaknesses. The mode is simple to compute, but there may be more than one. Further, if no data value appears more than once, the mode is entirely unhelpful. As for the mean and median, the main difference between these two measures is how each is affected by extreme values.

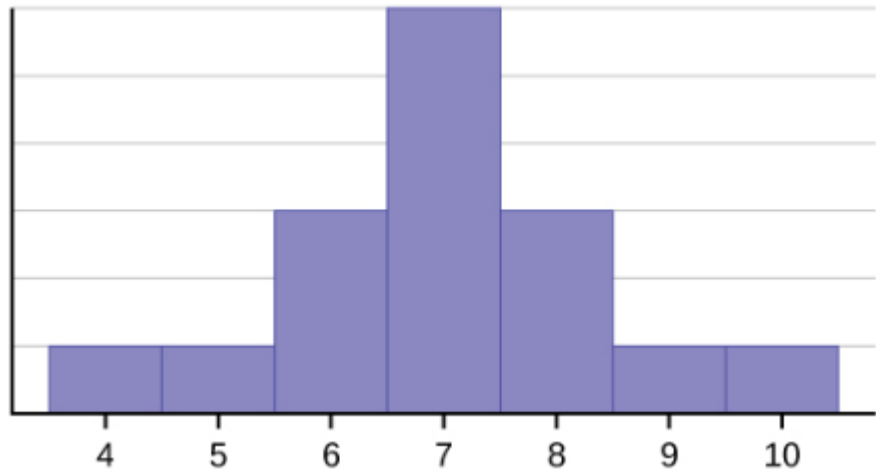
Consider this example: the mean and median of 1, 2, 3, 4, 5 are both 3. But what if the dataset is instead 1, 2, 3, 4, 10? The median is still 3, but the mean is now 4. What this example shows is that the mean is sensitive to extreme values, while the median isn't. This knowledge can help us decide which of the two is more relevant for a given dataset. If it is important that the really high or really low values are reflected in the measure of centrality, then the mean is the better option. If very high or low values are not important, however, then we should stick with the median.

The decision between mean and median only really matters if the data are skewed. If the data are symmetric, then the mean and median are going to be approximately equal, and the distinction between them is irrelevant. If the data are skewed, the mean is pulled in the direction of the skew (i.e., if the data are right-skewed, then the mean will be greater than the median; if the data are left-skewed, the opposite relation holds). The relationship between mean, mode, and median is given in the table below

Different Types of Distribution of the Data

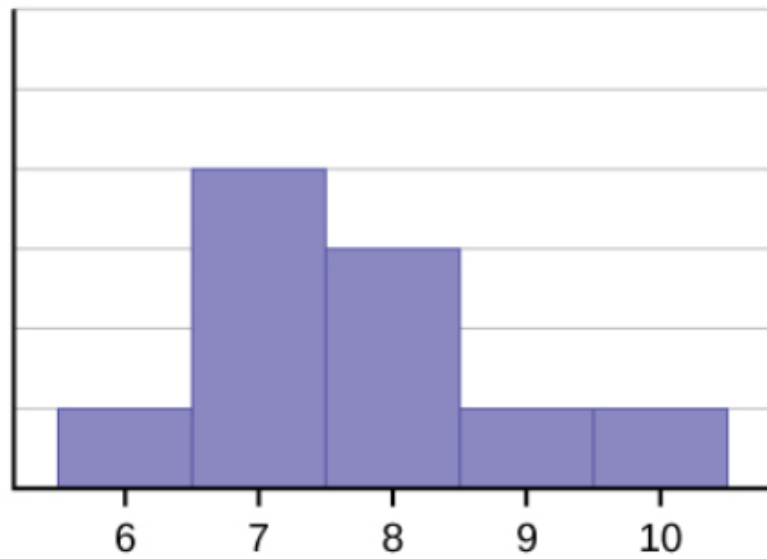
Symmetric Data

4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10



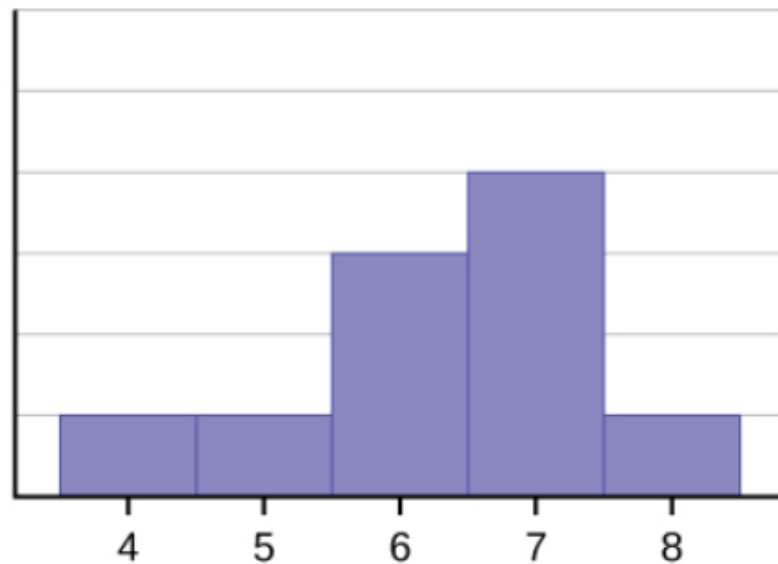
Skewed Right Data

6, 7, 7, 7, 7, 8, 8, 8, 9, 10



Skew Left Data

6, 7, 7, 7, 7, 8, 8, 8, 9, 10



Let us return now to our original household income data. To find the median, let's order the data

15 15 15 20 20 25 30 35 40 45 50 50 50 50 50

Since there are 15 data values, an odd number, the median will be the middle number, the  $\frac{15+1}{2} = 8^{\text{th}}$  data value. So  $8^{\text{th}}$  data value is 35.

The median income in this neighborhood sample is \$35 thousand. *Thus, half of the households' earned income is less than \$35,000 and the other half earned more than \$35,000.*

If we add the new neighbor with a \$2 million household income, then there will be 16 data values.

15 15 15 20 20 25 30 35 40 45 50 50 50 50 50 2000

The position of those two middle numbers are  $(\frac{16}{2} = 8)^{\text{th}}$  and  $(\frac{16}{2} + 1 = 9)^{\text{th}}$ . Which are 35 and 40.

$$\frac{(35 + 40)}{2} = 37.5$$

As we discovered in the last example, the median is \$35 thousand. Notice that the new neighbor did not affect the median in this case. **The median is less affected by outliers than the mean.**

### The "Billionaire in the Pub" Problem

To understand why the "average" can be misleading, imagine ten people sitting in a pub, each earning \$50,000 a year. Both the mean and median income are \$50,000.

If a billionaire walks in, the **group's mean** income suddenly jumps to nearly \$100 million. Looking at the average, you'd think everyone in the room was wealthy. However, the **median** remains \$50,000.

### Mean or Median: Which one is appropriate to use to describe data?

Use the median for **skewed data** or data with **extreme outliers**. Use the mean for **symmetric data**.

### Example 2.1.5: Which is Better Measure?

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 available at the end of this section.

### Your Turn 2.1.6: Comparing Mean and Median

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Your Turn 2.1.7: Mean or Median?

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Weighted Average and GPA

Have you considered how your grade point average (GPA) is calculated? Your business program requires the successful completion of many courses. Your grades in each course combine to determine your GPA; however, not every course necessarily has the same level of importance as measured by your course credits.

Perhaps your math course takes one hour daily, while your communications course is only delivered in one-hour sessions three times per week. Consequently, the college assigns the math course five credit hours and the communications course three credit hours. If you want an average, these different credit hours mean that the two courses do not share the same level of importance, and therefore, a simple average cannot be calculated.

In a weighted average, not all pieces of data share the same level of importance, nor do they occur with the same frequency. The data cannot represent a percent change or a series of numbers intended to be multiplied by each other. To calculate a weighted average, you require two components:

1. The data itself—you need the value for each piece of data.
2. The weight of the data—you need to know how important each piece of data is to the average. This is either an assigned value or a reflection of the number of times each piece of data occurs (the frequency).

To calculate a weighted average, add the products of the weights and data for the entire data set, and then divide this total by the total of the weights.

### Example 2.1.8: Find GPA

A mark transcript received by a student at a local college. The chart shows how the grade translates into a grade point.

Calculate the student's grade point average (GPA). Round your final answer to two decimals.

Answer available at the end of this section.

### Letter Grades and Credit Hours

| Course   | Letter Grade | Credit hours (Weight) |
|----------|--------------|-----------------------|
| ENC 1100 | B            | 4                     |
| MAC 2302 | A            | 5                     |
| PSY 1203 | B+           | 3                     |
| PHY 2017 | C            | 4                     |
| HUM 1048 | A+           | 3                     |
| MGF 1130 | D            | 4                     |

### Letter Grades and Corresponding GPAs

| Grade | Grade Point Value |
|-------|-------------------|
| A+    | 4.5               |
| A     | 4.0               |
| B+    | 3.5               |
| B     | 3.0               |
| C+    | 2.5               |
| C     | 2.0               |
| D     | 1.0               |
| F     | 0.0               |

Answer available at the end of this section.

### Your Turn 2.1.8: Find GPA

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example 2.1.9: Weighted Average

In your Mathematical Thinking class, your final grade is based on six categories. The categories are tests, labs, homework assignments, and a final exam. The final average for the course is the weighted average of scores earned in these categories with the following weights.

Table: Assignments and their Weights

| Assignments | Tests | Project | Homework | Final Exam | Quiz | Discussion |
|-------------|-------|---------|----------|------------|------|------------|
| Weights     | 50%   | 5%      | 15%      | 20%        | 5%   | 5%         |

Suppose you earned the following grades on average in each of the categories: 85% on tests, 0% on the project, 90% on homework assignments, 50% on the final exam, 80% on the quiz, and 100% on the discussion. Determine your weighted average in the course. Record the average below as a percentage accurate to two decimal places.

Answer available at the end of this section.

Answer available at the end of this section.

Answer available at the end of this section.

### Your Turn 2.1.9: Find Weighted Average

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Notes:

Disease in Childhood (paper) [openlearn.open.ac.uk]

[QT] Online Math Learning (opens in new window) [www.onlinemathlearning.com]

[FQ] Interquartile Range(opens in new window) [math.fandom.com]

## Answers

### Example 2.1.2: Find the Median

We start by listing the data in order:

2 4 5 5 6 7 7 8 8 10

Since there are 10 data values, an even number, there are two middle numbers. So, we find the mean of the two middle numbers. The position of those two middle numbers are  $(\frac{10}{2} = 5)^{\text{th}}$  and  $(\frac{10}{2} + 1 = 6)^{\text{th}}$ . Which are 6 and 7.

$$\frac{(6 + 7)}{2} = 6.5.$$

The median quiz score was 6.5.

*We can say that half of the quiz scores were lower than 6.5, and the other half were higher than 6.5.*

### Example 2.1.3: Finding Mode

1. Mode is 50,000.
2. The data set is bimodal because the scores 430 and 480 each occur twice.

### Example 2.1.5: Which is Better Measure?

There are 49 people who earn \$30,000. So we add \$30,000, 49 times. And one person who earns \$5,000,000.

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

$$\text{Median} = 30,000$$

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 2.1.8: Find GPA

### Example 2.1.8: Find GPA

The courses do not carry equal weights as they have different credit hours. Therefore, to calculate the GPA, you must find a weighted average.

Table: To Find the weighted Average

| Course   | Grade | Grade Point<br>(Data Value) | Credit hours (Weight) | Value x Weight        |
|----------|-------|-----------------------------|-----------------------|-----------------------|
| ENC 1100 | B     | 3.0                         | 4                     | $3 \times 4 = 12$     |
| MAC 2302 | A     | 4.0                         | 5                     | $4 \times 5 = 20$     |
| PSY 1203 | B+    | 3.5                         | 3                     | $3.5 \times 3 = 10.5$ |
| PHY 2017 | C     | 2.0                         | 4                     | $2 \times 4 = 8$      |
| HUM 1048 | A+    | 4.5                         | 3                     | $4.5 \times 3 = 13.5$ |
| MGF 1130 | D     | 1.0                         | 4                     | $1 \times 4 = 4$      |
| TOTAL    |       |                             | 23                    | 68                    |

$$\text{GPA} = \frac{68}{23} = 2.96$$

**Example 2.1.9: Weighted Average**

**Example 2.1.9: Weighted Average**

**Example 2.1.9: Weighted Average**

Multiply every group grade average by its associated weight.

Table: To find the weighted Average

| Course                   | Weight | Average Score<br>(Data Value) | Value x Weight              |
|--------------------------|--------|-------------------------------|-----------------------------|
| Tests                    | 0.50   | 85%                           | $0.50 \times 85\% = 42.5\%$ |
| Project                  | 0.05   | 0%                            | $0.05 \times 0\% = 0\%$     |
| Homework                 | 0.15   | 90%                           | $0.15 \times 90\% = 13.5\%$ |
| Final Exam               | 0.20   | 50%                           | $0.20 \times 50\% = 10\%$   |
| Quiz                     | 0.05   | 80%                           | $0.05 \times 80\% = 4\%$    |
| Discussion               | 0.05   | 100%                          | $0.05 \times 100\% = 5\%$   |
| Total (Weighted Average) | 1      |                               | 75%                         |

$$\text{Weight Average in the Course} = \frac{42.5\% + 0\% + 13.5\% + 10\% + 4\% + 5\%}{1} = 75\%$$

This page titled 2.1: Measures of Center of the Data is shared under a license and was authored, remixed, and/or curated by LibreTexts.

- 9.3: Measures of Central Tendency by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf).
- 8.3: Mean, Median and Mode by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/contemporary-mathematics>.
- 9.4: Measures of Variation by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf).

## 2.2: Measures of Spread of the Data

### Measure of Dispersion

A **measure of dispersion** is a statistic that describes how spread out or scattered the data values are in a data set.

- **Range** – difference between the maximum and minimum values
- **Variance** – average squared distance from the mean
- **Standard deviation** – typical distance from the mean
- **Interquartile range (IQR)** – spread of the middle 50% of the data (We study this in next section)

#### Standard Deviation

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

Think of two basketball players who both average 20 **points per game**.

- **Player A** scores exactly 20, 19, 21, and 20. Their standard deviation is **low** because their scores are very close to the average.
- **Player B** scores 5, 35, 10, and 30. Their standard deviation is **high** because their scores are volatile and spread far from the average.

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

Consider these three sets of quiz scores:

Section A:

5 5 5 5 5 5 5 5 5

Section B:

0 0 0 0 0 10 10 10 10 10

Section C:

4 4 4 5 5 5 5 6 6 6

All three of these sets of data have a mean of 5 and a median of 5, yet the sets of scores are clearly quite different. In section A, everyone had the same score; in section B, half the class got no points and the other half got a perfect score, assuming this was a 10-point quiz. Section C was not as consistent as Section A, but not as widely varied as Section B.

In addition to the mean and median, which are measures of the “typical” or “middle” value, we also need a measure of how “spread out” or varied each data set is.

There are several ways to measure this “spread” of the data, as we mentioned above, these are called measures of dispersion. The first is the simplest and is called the **range**.

#### Definition: Range

The range is the difference between the maximum and minimum values of a data set.

Range = Largest data value – smallest data value

### Example 2.2.1: Find Range

Using the quiz scores from above, find the range of each section.

*Answer available at the end of this section.*

In the last example, the range appears to reveal how spread out the data is. However, suppose we add a fourth section, Section D, with scores

Section D: 0 5 5 5 5 5 5 5 5 10

This section also has a mean and median of 5. The range is 10, yet this data set is quite different from **Section B**. To better illuminate the differences, we'll have to turn to more sophisticated measures of variation, such as the standard deviation.

### Sample and Population Standard Deviation and a Symbol to Represent Them

The standard deviation of the entire population is called the population standard deviation. The standard deviation of a sample of the population is called the sample standard deviation.

Sample standard deviation: Symbol to represent is  $s$

Population standard deviation: Symbol to represent is  $\sigma$  (Read as "sigma")

### Formula and Steps to Compute Standard Deviation and Variance

1. Find the deviation of each data point from the mean. In other words, subtract the mean from the data value.
2. Square each deviation.
3. Add the squared deviations.
4. Divide by  $n$ , the number of data values, if the data represents a whole population; divide by  $n-1$  if the data is from a sample.
5. Compute the square root of the result. (This result is the standard deviation.)

The above steps are summarized in the given formula

Sample standard deviation  $s$

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

Population standard deviation  $\sigma$

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

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

**Variance**

Variance is the square of the standard deviation. So sample variance is  $s^2$  and population variance is  $\sigma^2$

### Example 2.2.2: Calculate Standard Deviation

For the given data for section D quiz score:

Section D: 0 5 5 5 5 5 5 5 5 10

Find the deviation of the data from the mean and compute the standard deviation and variance of the data.

Answer available at the end of this section.

### Your Turn 2.2.2: Compare Standard Deviation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



S

### Example 2.2.3: Finding Standard Deviation

Computing the standard deviation for Section B above, we first calculate that the mean is 5. Using a table can help keep track of your computations for the standard deviation:

| Data Value | Deviation: Data Value - Mean | (Deviation) <sup>2</sup> |
|------------|------------------------------|--------------------------|
| 0          | $0 - 5 = -5$                 | $(-5)^2 = 25$            |
| 0          | $0 - 5 = -5$                 | $(-5)^2 = 25$            |
| 0          | $0 - 5 = -5$                 | $(-5)^2 = 25$            |
| 0          | $0 - 5 = -5$                 | $(-5)^2 = 25$            |
| 0          | $0 - 5 = -5$                 | $(-5)^2 = 25$            |
| 10         | $10 - 5 = 5$                 | $(5)^2 = 25$             |
| 10         | $10 - 5 = 5$                 | $(5)^2 = 25$             |
| 10         | $10 - 5 = 5$                 | $(5)^2 = 25$             |
| 10         | $10 - 5 = 5$                 | $(5)^2 = 25$             |
| 10         | $10 - 5 = 5$                 | $(5)^2 = 25$             |

Assuming this data represents a sample, we will add the squared deviations, divide by  $10 - 1 = 9$ , the number of data values, and compute the square root:

$$s = \sqrt{\frac{25 + 25 + 25 + 25 + 25 + 25 + 25 + 25 + 25 + 25}{10 - 1}} = \sqrt{\frac{250}{9}} = 5.27$$

Notice that the standard deviation of this data set is much larger than that of section D, since the data in this set is more spread out.

## Comparing the Standard Deviation of Four Data Sets

| Section A: 5 5 5 5 5 5 5 5 5 5      | Standard deviation: 0    |
|-------------------------------------|--------------------------|
| Section B: 0 0 0 0 0 10 10 10 10 10 | Standard deviation: 5.27 |
| Section C: 4 4 4 5 5 5 5 6 6 6      | Standard deviation: 0.8  |
| Section D: 0 5 5 5 5 5 5 5 5 10     | Standard deviation: 2.4  |

### Your Turn 2.2.3: Finding Standard Deviation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Interpreting the Standard Deviation

The standard deviation,  $s$  or  $\sigma$ , 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  $\sigma$  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.

As we know standard deviation measures the "typical" distance between each data point and the mean. A low standard deviation indicates that the values are clustered closely around the average, suggesting high consistency or precision. Conversely, a high standard deviation shows that the data is spread out over a wider range, indicating more variability or "noise" within the set.

For example, imagine two archery students, Alex and Sam, who both have an average distance of 5 inches from the bullseye. If Alex has a standard deviation of 1 inch, his arrows are consistently near the center. If Sam has a standard deviation of 10 inches, his arrows are scattered wildly across the target; despite having the same average as Alex, Sam's performance is much

less predictable.

### Your Turn 2.2.4: Interpreting Standard Deviation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Rule of Thumb to Estimate Standard Deviation

The Range Rule of Thumb is generally used to estimate population standard deviation. It acts as a quick heuristic, best suited for normal distributions with a sample size of approximately 15 to 35, although it is often used loosely for general data sets.

$$\sigma \approx \frac{\text{Range}}{4}$$

Note: If the rule-of-thumb estimate differs from the actual standard deviation by more than about 20 it is generally considered not a good estimate.

## The Empirical Rule

We start by examining a specific set of data. Table 2.2.1 shows the heights in inches of 100 randomly selected adult men. A relative frequency histogram for the data is shown in Figure 2.2.1. The mean and standard deviation of the data, rounded to two decimal places, are  $M = 69.92$  and  $SD = 1.70$ .

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

Table 2.2.1: Height of Men

If we go through the data and count the number of observations that are within one standard deviation of the mean, that is, that are between  $69.92 - 1.70 = 68.22$  and  $69.92 + 1.70 = 71.62$  inches, there are 69 of them.

If we count the number of observations that are within two standard deviations of the mean, that is, that are between  $69.92 - 2(1.70) = 66.52$  and  $69.92 + 2(1.70) = 73.32$  inches, there are 95 of them.

All of the measurements are within three standard deviations of the mean, that is, between  $69.92 - 3(1.70) = 64.822$  and  $69.92 + 3(1.70) = 75.02$  inches.

These tallies are not coincidences but rather align with the following result, which has been found to be widely applicable.

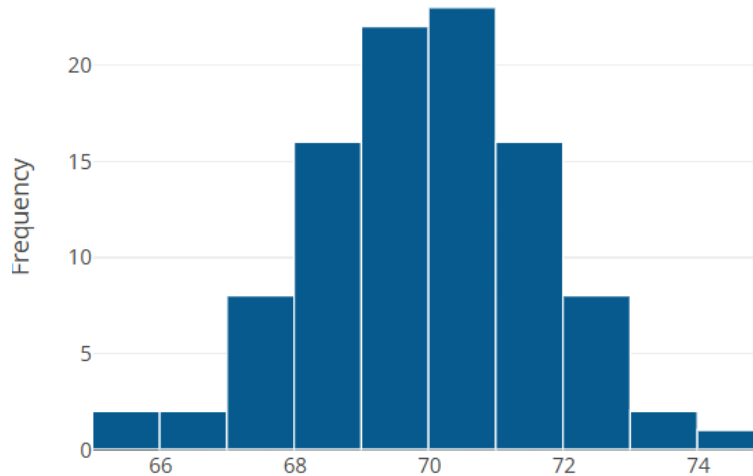


Figure 2.2.2: Normal (Bell Shaped) Distribution

### Empirical Rule

The Empirical Rule (also called the 68–95–99.7 rule describes how data are distributed in a normal (bell-shaped) distribution

1. approximately 68% of the data lie between one standard deviation mean, i.e., within  $(M - 1 \times SD, M + 1 \times SD)$ .
2. approximately 95% of the data lie between one standard deviation mean, i.e., within  $(M - 2 \times SD, M + 2 \times SD)$ .
3. approximately 99.7% of the data lie between one standard deviation mean, i.e., within  $(M - 3 \times SD, M + 3 \times SD)$ .

Where M is the mean, and SD is the standard deviation.

Two key points in regard to the Empirical Rule are that the data distribution must be approximately bell-shaped and that the percentages are only approximately true. The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule.

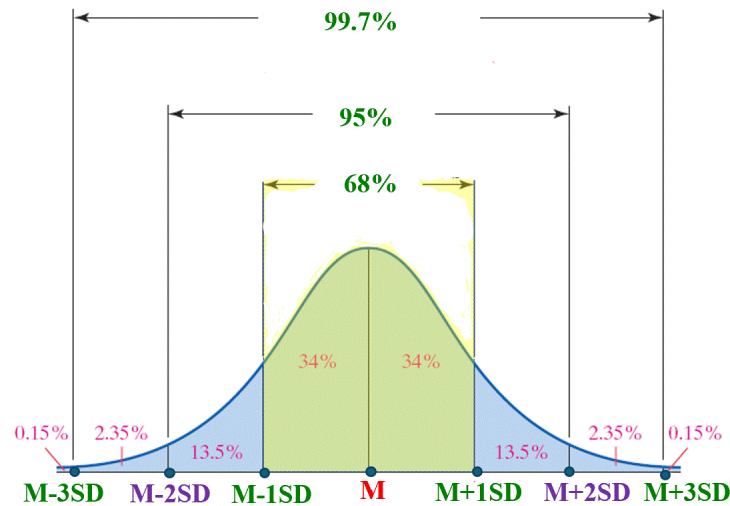


Figure 2.2.3: M as a Mean and SD as a Standard Deviation

We see this with the example of the heights of the men: the Empirical Rule suggested 68 observations between 68.22 and 71.62 inches, but we counted 69.

### Example 2.2.5: Using Empirical Rule

Scores on IQ tests have a bell-shaped distribution with a mean of 100 and a standard deviation of 10. Discuss what the Empirical Rule is after sketching a normal distribution curve. What is the percentile of IQ scores of 110 and 90?

*Answer available at the end of this section.*

### Your Turn 2.2.5: Use Empirical Rule

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 2.2.6: Using Empirical Rule

1. About what percentage of all such men are between 68.2 and 71 inches tall?
2. What interval centered on the mean should contain about 95% of all such men?
3. What percentage of all such men are less than 66.8 inches in tall?

Answer available at the end of this section.

### Your Turn 2.2.6: Use Empirical Rule

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



Consider the following scenario:

Suppose we are studying how long customers wait in line at the checkout counters at supermarkets *A* and *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 there is greater variation in its wait times. Overall, wait times at supermarket *B* are more spread out from the average; those at supermarket *A* are more concentrated around the average.

Suppose that Rosa and Bina both shop at supermarket *A*.

Rosa waits at the checkout counter for seven minutes.

Bina 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.

### Below the Average or Above the Average?

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.)

### Meaning of Intervals in Empirical Rules

1. less than 3 standard deviations below the mean: Outliers
2. Between 2 and 3 standard deviations below the mean: significantly below the average
3. Between 1 and 2 standard deviations below the mean: below the average
4. Within 1 standard deviations from the mean: average

5. More than 3 standard deviations below the mean: Outliers
6. Between 2 and 3 standard deviations above the mean: significantly above the average
7. Between 1 and 2 standard deviations above the mean: above the average.

### Your Turn 2.2.7: Use Empirical Rule

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Answers

### Example 2.2.1: Find Range

For section A, the range is 0 since both maximum and minimum are 5 and  $5 - 5 = 0$

For section B, the range is 10 since  $10 - 0 = 10$

For section C, the range is 2 since  $6 - 4 = 2$

### Example 2.2.2: Calculate Standard Deviation

Mean is 5 for the given data set.

Table: Calculate Standard Deviation of the Data

| Data Value | Deviation: Data Value - Mean | (Deviation) <sup>2</sup> |
|------------|------------------------------|--------------------------|
| 0          | $0 - 5 = -5$                 | $(-5)^2 = 25$            |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 5          | $5 - 5 = 0$                  | $(0)^2 = 0$              |
| 10         | $10 - 5 = 5$                 | $(5)^2 = 25$             |

We then add the squared deviations up to get

$$25 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 25 = 50$$

Variance can be a useful statistical concept, but note that the units of variance in this instance would be points-squared since we squared all of the deviations. What are points-squared? Good question. We would rather deal with the units we started with (points in this case), so to convert back, we take the square root and get:

$$\text{Population Standard Deviation } \sigma = \sqrt{\frac{50}{10}} = \sqrt{5} \approx 2.2$$

$$\text{Sample Standard Deviation } s = \sqrt{\frac{50}{9}} \approx 2.4$$

$$\text{Population Variance } \sigma^2 = 5$$

$$\text{Sample Variance } s^2 = 5.56$$

If we are unsure whether the data set represents a sample or a population, we typically assume it is a sample.

### Example 2.2.5: Using Empirical Rule

A sketch of the IQ distribution is given in Figure 2.2.3. The Empirical Rule states that

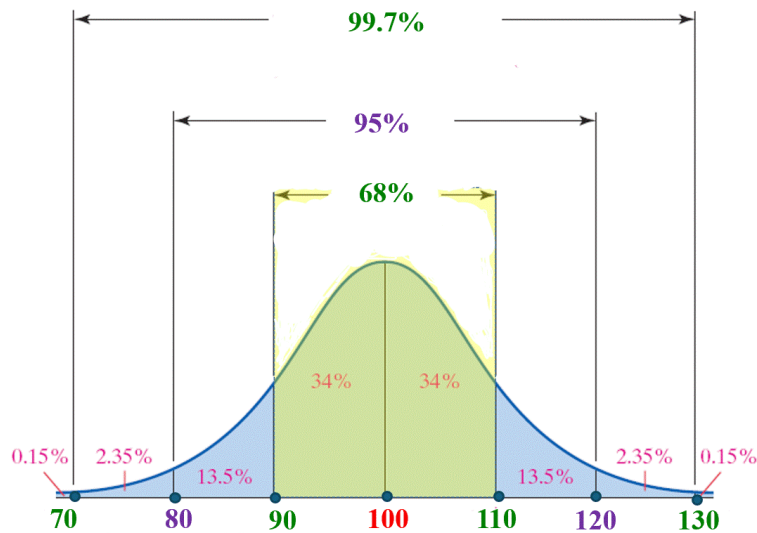


Figure 2.2.4: Distribution of IQ Scores

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

Since 84% data are less than or equal to 110, we conclude that the IQ score 110 is the 84<sup>th</sup> percentile. Similarly, Since 16% data are less than or equal to 90, we conclude that the IQ score 90 is the 16<sup>th</sup> percentile.

#### Example 2.2.6: Using Empirical Rule

1. A sketch of the distribution of heights is given in Figure 2.2.3. Since the interval from 68.2 to 71.0 has endpoints  $\text{Mean} - 1 \times \text{SD}$  and  $\text{Mean} + 1 \times \text{SD}$ , by the Empirical Rule about 68% of all 18-year-old males should have heights in this range
2. By the Empirical Rule the shortest such interval has endpoints

$$\text{Mean} - 2 \times \text{SD} = 69.6 - 2(1.4) = 66.8$$

$$\text{Mean} + 2 \times \text{SD} = 69.6 + 2(1.4) = 72.4$$

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

3.  $0.15\% + 2.35\% = 2.5\%$

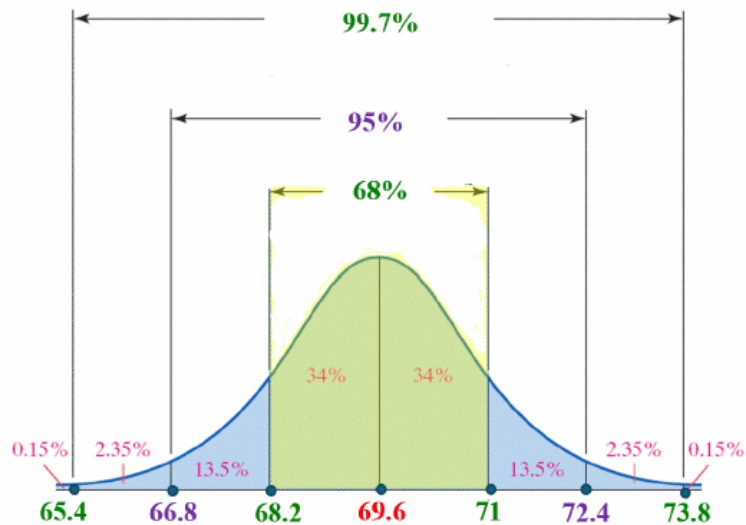


Figure 2.2.5: Distribution of Heights

1.

---

This page titled [2.2: Measures of Spread of the Data](#) is shared under a 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: Measures of Variation](#) by [Darlene Diaz](#) is licensed [CC BY-NC-SA 4.0](#). Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH%20105/Math%20For%20Liberal%20Art%20Students%20(2017).pdf).
- [Taylor Expansion II](#) has no license indicated.
- [Taylor Expansion](#) has no license indicated.

## 2.3: Measures of Location of the Data

### Quartiles

Where standard deviation is a measure of variation based on the mean, **quartiles** are based on the median. Quartiles are values that divide a dataset into four equal parts after the data has been arranged in order from smallest to largest.

#### Definition: Quartiles and Interpretation

Quartiles are values that divide a dataset into four equal parts, with each part containing 25% of the data points

The first quartile  $Q_1$  is the value so that 25% of the data values are below or equal to it. The third quartile  $Q_3$  is the value so that 75% of the data values are less than or equal to it.

You may have guessed that the second quartile is the same as the median,  $Q_1$ . The median is the value such that 50% of the data values are below or equal to it.

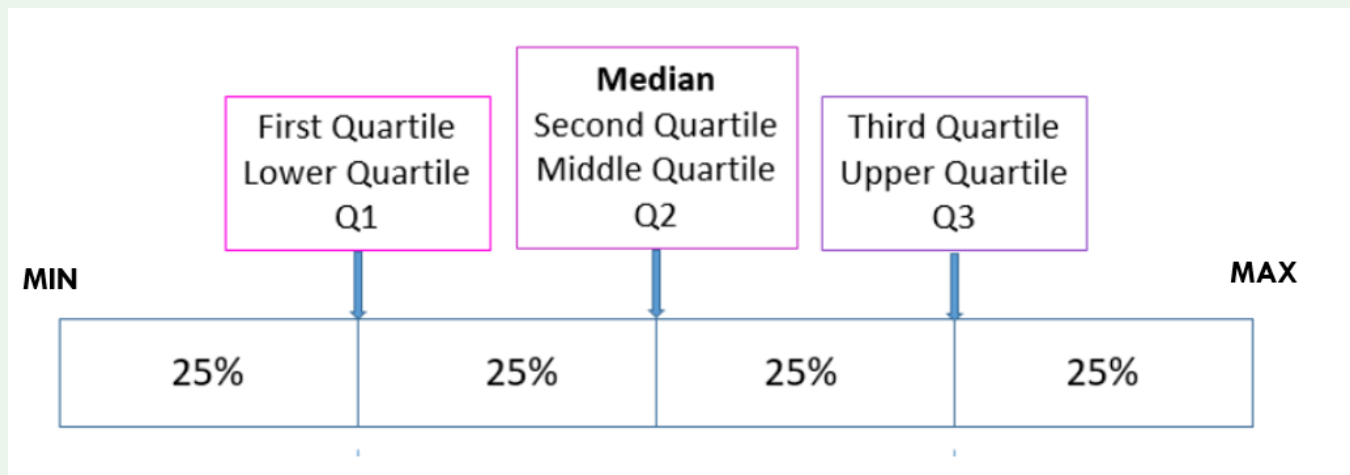


Figure 2.3.1: Quartiles (QT)

Quartiles divide the data into quarters, as we see above Figure 2.3.1.

The following are the **interpretations** of the quartiles.

- 25% of the data is between the minimum and  $Q_1$ .
- 25% of the data is between  $Q_1$  and the median ( $Q_2$ ).
- 50% of the data is between  $Q_1$  and  $Q_3$ .
- 75% of the data is above  $Q_3$ .

Notations:

First Quartile  $\equiv$  Lower Quartile  $\equiv Q_1$ : ( $Q_1$ ) is also called 25<sup>th</sup> percentile.

Second Quartile  $\equiv$  Median  $\equiv Q_2 \equiv$  Middle Quartile: ( $Q_2$ ) is also called 50<sup>th</sup> percentile.

Third Quartile  $\equiv$  Upper Quartile  $\equiv Q_3$ ; ( $Q_3$ ) is also called 75<sup>th</sup> percentile.

### How to Find $Q_1$ and $Q_3$

1. Begin by ordering the data from smallest to largest.
2. Find the median of the data values, which is the middle quartile.
3. Find the median of the data values that are smaller than the median. Those data values are referred to as the lower half. Which is the lower quartile,  $Q_1$ .
4. Find the median of the data values that are larger than the median. Those data values are referred to as the upper half. Which is the upper quartile,  $Q_3$ .

Let's find the quartiles for the ordered data given below

45 47 52 52 53 55 56 58 62 80

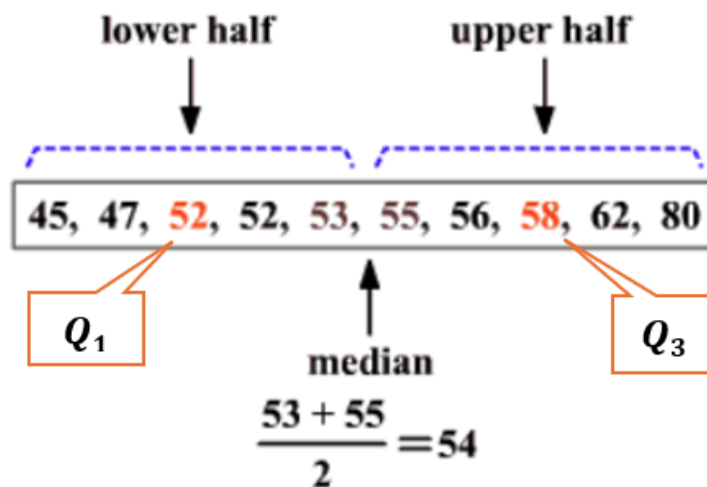


Figure 2.3.2: Finding Quartiles [FQ]

Let's look at some examples. We can also calculate the five-number summary using calculators or scientific software such as Excel, Minitab, or R. However, in this course, we only get our feet wet with statistics, so we can calculate these values quickly by hand.

### Example 2.3.1: Finding Quartiles

Suppose we have measured 9 females and their heights (in inches), sorted from smallest to largest, are given below. Find the Quartiles.

59 60 62 64 66 67 69 70 72

Answer available at the end of this section.

### Your Turn 2.3.1: Find Quartiles

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Fences, IQR, and Outliers

### Interquartile Range (IQR)

The interquartile range (IQR) is a measure of how spread out the middle 50% of a dataset is. It is calculated as the difference between the upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ ).

$$IQR = Q_3 - Q_1$$

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

### Fences and Outliers and $IQR \times 1.5$ Rule

An outlier is a data point that is significantly different from the other data points.

**Fences** are boundaries used to identify potential outliers in a dataset. They are based on quartiles and the interquartile range (IQR).

Lower fence (LF) and upper fence are calculated using the following formula

$$LF = Q_1 - (1.5)(IQR)$$

$$UF = Q_3 + (1.5)(IQR)$$

### Example 2.3.2: Box Plot with Outliers

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 available at the end of this section.

### Your Turn 2.3.2: Find Quartiles

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Five Number Summary and Box Plot

While quartiles are not a single-number summary of variation like standard deviation, they are used in conjunction with the median, minimum, and maximum values to form a **5-number summary** of the data.

### Definition: Five-Number Summary

The five-number summary takes this form

Minimum,  $Q_1$ , Median,  $Q_3$ , Maximum

How to find the quartiles in the data set?

The 5-number summary combines the first and third quartiles with the minimum, median, and maximum values.

In the example with a sample of 9 females, the median is 66, the minimum is 59, and the maximum is 72. Hence, the 5-number summary is:

59 61 66 69.5 72

Note that the five-number summary divides the data into four intervals, each of which will contain about 25% of the data. For visualizing data, there is a graphical representation of a five-number summary called a box plot, or box-and-whisker graph.

For visualizing data, there is a graphical representation of a five-number summary called a **box plot**, or box-and-whisker graph.

### Definition: Box Plot

A **box plot** is a graphical representation of a five-number summary.

To create a box plot, a number line is first drawn with equidistant tick marks. A box is drawn from the first quartile to the third quartile, and a line is drawn through the box at the median. "Whiskers" are extended out to the minimum and maximum values.

### Example 2.3.3: Box Plot

What is the box plot of the data in example 2.3.1 and 2.3.2 ?

Answer available at the end of this section.

### Your Turn 2.3.3: Find Quartiles and Box Plot

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



Box plots are particularly useful for comparing data from two populations or samples. In fact, when comparing two samples, it is always preferable to use box plots.

### Example 2.3.4: Reading Box Plot

The box plot of service times for two fast-food restaurants is shown below.

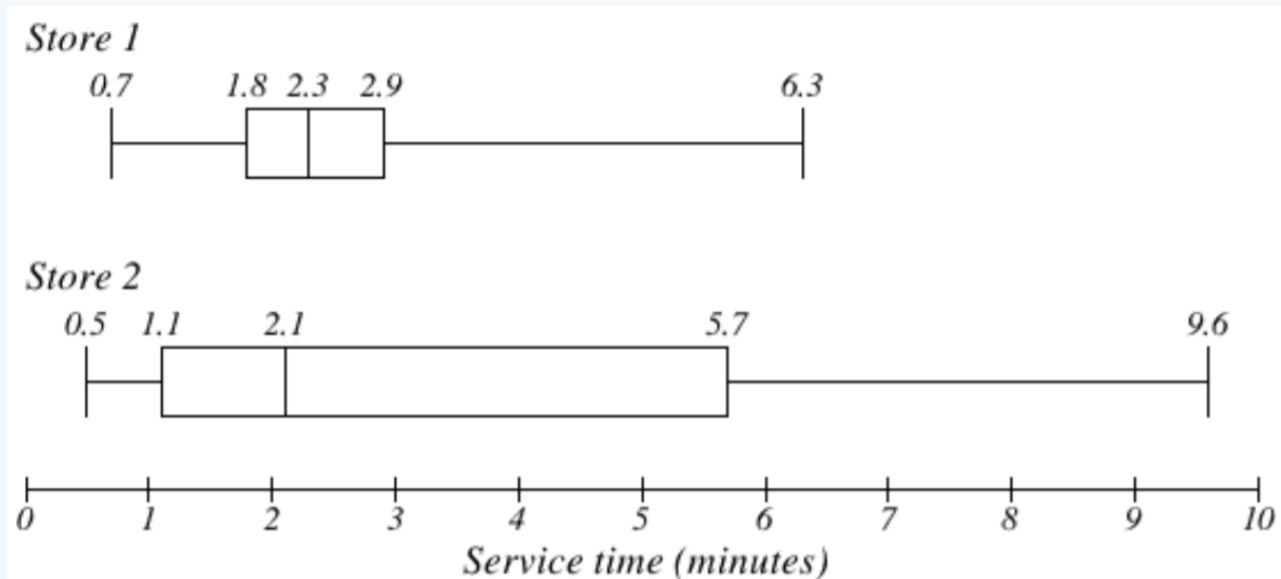


Figure 2.3.4: Side by Side Two Box Plot

1. Which fast-food restaurant has a smaller median service time but a more spread-out distribution of service times?
2. In Store 1, what percentage of customers were served within 2.9 minutes?
3. In Store 1, 75% of customers were served in less than what amount of time?
4. In Store 2, what percentage of customers waited more than 5.7 minutes?
5. If Alex waited 4 **minutes**, and this wait time is longer than that of 90% of customers, was he in Store 1 or **Store 2**?

Answer available at the end of this section.

### Your Turn 2.3.4: Reading Box Plot

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Percentiles

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

As we know, quartiles in the previous section. Now we will introduce the percentile.

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

A **percentile** is a measure of ranking. It represents a location measurement of a data value to the rest of the values. Many standardized tests give the results as a percentile. Doctors also use percentiles to track a child's growth.

### Interpretation: Percentile of Data

The  $k^{\text{th}}$  percentile is the data value that contains  $k\%$  of the data at or below it.

$Q_1$  is 25<sup>th</sup> percentile. That is, 25% of the data is less than or equal to  $Q_1$ .

$Q_2$  is 50<sup>th</sup> percentile. That is, 50% of the data is less than or equal to  $Q_2$ .

$Q_3$  is 75<sup>th</sup> percentile. That is, 75% of the data is less than or equal to  $Q_3$ .

The  $P^{\text{th}}$  percentile cuts the data set in two so that approximately  $P\%$  of the data lie below or equal to it and  $(100 - P)\%$  of the data lie above it. In particular, the three percentiles that cut the data into fourths are called the quartiles of a data set. The quartiles

are the three numbers  $Q_1$ ,  $Q_2$ ,  $Q_3$  that divide the data set approximately into fourths.

### What Do You Mean by Data Value $x$ is in $P^{th}$ percentile?

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

### Example 2.3.5: Interpreting Percentile

- What does a score of the 90<sup>th</sup> percentile mean? If a sample consists of 800 test scores, how many of them would be at or below the 90<sup>th</sup> percentile?
- How many scores are above the 90<sup>th</sup> percentile?

*Answer available at the end of this section.*

### Example 2.3.6: Percentile Vs Score

If the test was out of 100 points and you scored at the 80<sup>th</sup> percentile, what was your score on the test?

*Answer available at the end of this section.*

### Your Turn 2.3.6: Interpret Percentile

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## 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.

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.

### Example 2.3.7: Finding k-th Percentile

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 70<sup>th</sup> percentile.
- Find the 83<sup>rd</sup> percentile.

Answer available at the end of this section.

### Your Turn 2.3.7: Find kth Percentile

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## $z$ -Scores and Normal Distribution

You have probably seen a "bell-shaped curve" before. It turns out that many data sets, particularly when the sample size is sufficiently large, are distributed symmetrically, with most data values around the center (mean) and few outliers, resulting in a bell-shaped curve. This is called a "normal distribution." For instance, the weights of all the people in a certain country are likely distributed normally, and so are their heights. Normal distributions can describe all sorts of data involving people's cholesterol levels, blood pressure, duration of pregnancy, IQ (intelligence quotient) scores, credit scores, and many more. For instance, the pregnancy (gestation) period for humans is known to have an average of 266 days (38 weeks) with a standard deviation of 16 days. The average IQ score is defined as 100 with a standard deviation of 15. Your SAT and other standardized test scores are defined in similar ways.

But what does it mean, and how does this information help us?

When data values are symmetric, given their mean and standard deviation, we can then standardize the normal distribution and calculate various percentages related to scores/values. For example, given that the IQ scores have a mean of 100 and a standard deviation of 15, one can conclude that only about 2.3% of the population has an IQ of 130 or higher.

It turns out that there is a simple way to "convert" any normal distribution to the standard one. It involves what is known as the " $z$ -score."

Another way to locate a particular observation  $x$  in a data set is to compute its distance from the mean in units of standard deviation. The  $z$ -score indicates how many standard deviations an individual observation  $x$  is from the center of the data set, its mean. It is used on distributions that have been *standardized*, which allows us to better understand its properties. If  $z$  is negative, then  $x$  is below average. If  $z$  is 0 then  $x$  is equal to the average. If  $z$  is positive, then  $x$  is above the average.

### What is a $z$ -score?

A  **$z$ -score** tells you how many standard deviations a value is away from the mean (average)

- Positive  $Z$ -score: The value is above average.
- Negative  $Z$ -score: The value is below average.

- Z-score of 0: The value is exactly the average.

The following formula is used to calculate the z-score ( $z$ ) of a data value ( $x$ ) when the mean, standard deviation, and data value ( $x$ ) are provided.

#### Formula: Finding Z-score from Data

$$z - \text{score of data} = \frac{\text{data value}(x) - \text{mean (M)}}{\text{standard deviation (SD)}}$$

Sometime  $z$  score of data value  $x$  is denoted by  $z_x$  or simply  $z$ .

#### Another notation:

Mean:  $\mu$  or  $\bar{x}$

Standard Deviation:  $\sigma$  or  $s$

#### Example 2.3.8: Find z-Score

On a nationwide math test, the mean was 65 and the standard deviation was 10. If Robert scored 81, what was his  $z$ -score?

*Answer available at the end of this section.*

#### Your Turn 2.3.8: Find z-score

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



The following formula is used to calculate the data value ( $x$ ) given the mean, standard deviation, and  $z$ -score.

#### Formula: Find Data Value from Z-Score

$$\text{data} = \text{mean} + z - \text{data value} \times \text{standard deviation}$$

#### Example 2.3.9: Find the Data from the z-score

A math test has mean 70 and standard deviation 8. A student has a  $z$ -score of 1.5 on this test. What is the student's actual score?

Answer available at the end of this section.

### Your Turn 2.3.9: Find the Date from the z-score

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example 2.3.10: Compare Using z-scores

A server at a restaurant tracks their weekly tips earned and hours worked. Over several weeks, the average weekly tips earned is \$500 with a standard deviation of \$100. This week, the server earned \$650 in tips. Meanwhile, the average number of hours worked per week is 30 hours with a standard deviation of 5 hours, and this week, the server worked 40 hours.

Using z-scores, determine whether the server's weekly tips or weekly hours worked deviate more from the average, indicating where their performance was stronger relative to past trends.

Answer available at the end of this section.

### Your Turn 2.3.10: Who is Heavier?

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Notes:

[QT] [Online Math Learning](http://www.onlinemathlearning.com) (opens in new window) [www.onlinemathlearning.com]

[FQ] [Interquartile Range](http://math.fandom.com)(opens in new window) [math.fandom.com]

## Answers

### Example 2.3.1: Finding Quartiles

Median is 66 inches. We can say that 50% or  $\frac{1}{2}$  of females are shorter than or equal to 66 inches, and the other 50% or  $\frac{1}{2}$  are taller than 65 inches. Data values that are smaller than the medians are

$$\begin{array}{cccc} 59 & 60 & 62 & 64 \\ \text{Lower Quartile} = \frac{(60 + 62)}{2} = 61 \end{array}$$

We can say that 25% or  $\frac{1}{4}$  of females are shorter than or equal to 61 inches, and the other 75% or  $\frac{3}{4}$  are taller than 61 inches. Data values that are larger than the medians are

$$\begin{array}{cccc} 67 & 69 & 70 & 72 \\ \text{Upper Quartile} = \frac{(69 + 70)}{2} = 69.5 \end{array}$$

We can say that 75% or  $\frac{3}{4}$  of females are shorter than or equal to 69.5 inches, and the other 25% or  $\frac{1}{4}$  are taller than 69.5 inches.

### Example 2.3.2: Box Plot with Outliers

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

$$Q_1 = \$40,500$$

$$Q_3 = \$69,000$$

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

$$\begin{aligned} LF &= Q_1 - (1.5)(IQR) \\ &= \$40,500 - (1.5)(\$28,500) \\ &= \$40,500 - \$42,750 \\ &= -\$2,250 \end{aligned}$$

$$\begin{aligned} UF &= Q_3 + (1.5)(IQR) \\ &= \$69,000 + (1.5)(\$28,500) \\ &= \$69,000 + \$42,750 \\ &= \$111,750 \end{aligned}$$

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

### Example 2.3.3: Box Plot

The box plot below is based on the 5-number summary from 2.3.1, which is given below.

Five number summary: 59    61    66    69.5    72

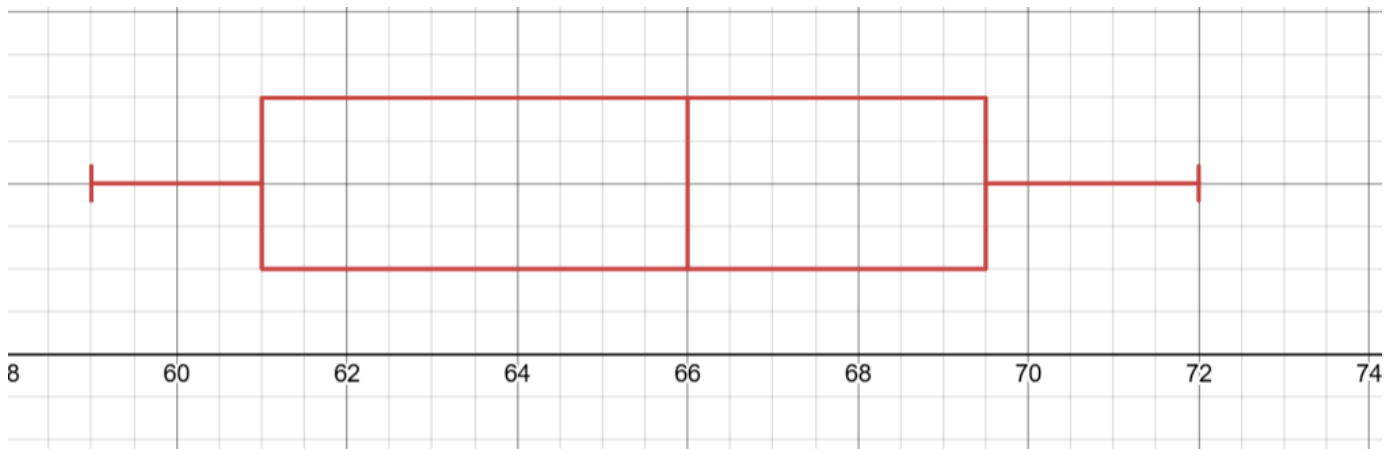


Figure 2.3.3: Box Plot of female heights (No outliers)

The box plot below is based on the 5-number summary from 2.3.2, which is given below.

Five number summary: 28,000    40,500    54,000    69,000    120,000

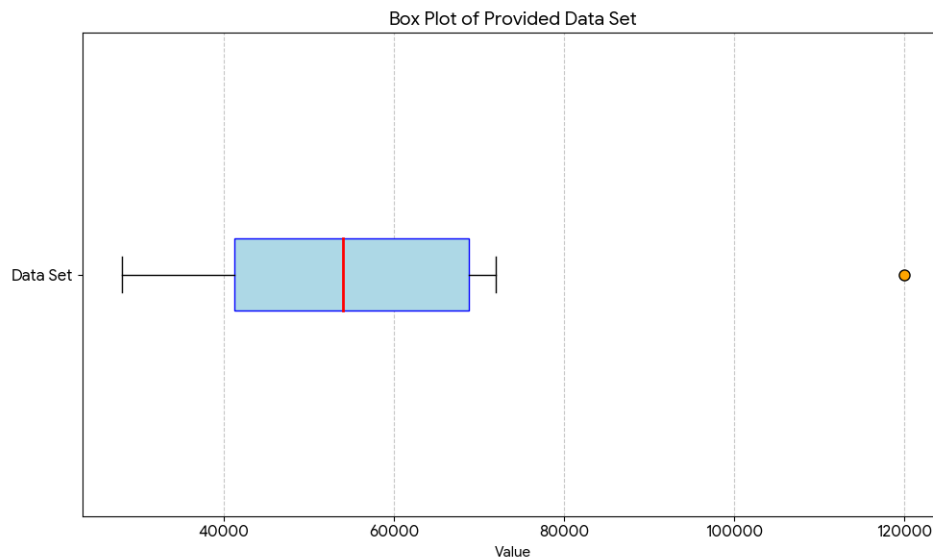


Figure 2.3.1: Box plot of salaires with outliers

### Example 2.3.4: Reading Box Plot

1. While Store 2 had a slightly shorter median service time (2.1 minutes vs. 2.3 minutes), Store 2 is less consistent, with a wider spread of the data.
2. In Store 1, 75% of customers were served within 2.9 minutes.
3. In Store 2, 75% of customers were served within less than 5.7 minutes.

4. 25%
5. If Alex waited 4 minutes, and this wait time is longer than that of 90% of customers, he is in Store 1.

#### Example 2.3.5: Interpreting Percentile

1. This means that 90% of the scores were at or below this score. (A person did better than 90% of the test takers.)

$$0.90 \times 800 = 720$$

So 720 score would be below or at 90<sup>th</sup> percentile.

2.  $800 - 720 = 80$  of the scores were above 90<sup>th</sup> percentile.

#### Example 2.3.6: Percentile Vs Score

You don't know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were low, you could have still failed the test. On the other hand, if many of the scores were high, you could have gotten.

#### Example 2.3.7: Finding k-th Percentile

$$k = 70$$

$i$  = ranking or position of a data value

$$n = 29$$

The ranking position of the data value is

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

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. So  $P_{70} = 64$

$$k = 83^{\text{rd}} \text{ percentile}$$

$i$  = ranking or position of a data value

$$n = 29$$

The ranking position of the data value is

$$i = \frac{k}{100}(n + 1) = \frac{83}{100}(29 + 1) = 24.9$$

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 is  $\frac{71+72}{2} = 71.5$  The 83rd percentile is 71.5 years. So  $P_{83} = 71.5$

#### Example 2.3.8: Find z-Score

$$\begin{aligned} z &= \frac{x - M}{SD} \\ &= \frac{81 - 65}{10} \\ &= \frac{16}{10} \\ &= 1.6 \end{aligned}$$

This means Robert's score was 1.6 standard deviations above the mean.

#### Example 2.3.9: Find the Data from the z-score

Here, we need to find the student's score (data value), denoted as  $x$ .

$$\begin{aligned} \text{data}(x) &= M + z - \text{score} \times SD \\ &= 70 + 1.5 \times 8 \\ &= 82 \end{aligned}$$

This means Robert's score was 82 corresponding to z-score 1.5.

#### Example 2.3.10: Compare Using z-scores

Use the z-score formula.

z-score for weekly tips:

$$z = \frac{650 - 500}{100} = \frac{150}{100} = 1.5$$

z-score for weekly hours worked:

$$z = \frac{40 - 30}{5} = \frac{10}{5} = 2.0$$

The z-score for hours worked 92 is higher than the z-score for tips earned 1.5.

This means that, relative to their average, the server worked significantly more hours than usual for the amount of tips they earned. Therefore, their hours worked deviated more from the average than their tips.

---

This page titled [2.3: Measures of Location of the Data](#) is shared under a license and was authored, remixed, and/or curated by LibreTexts.

- [9.4: Measures of Variation](#) by [Darlene Diaz](#) is licensed [CC BY-NC-SA 4.0](#). Original source: [https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students \(2017\).pdf](https://www.sccollege.edu/OER/Documents/MATH%205/Math%20For%20Liberal%20Art%20Students%20(2017).pdf).
- [Taylor Expansion](#) has no license indicated.
- [9.2: Solve Quadratic Equations Using the Square Root Property](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.

## 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.

### 3.1: Scatter Diagrams and Correlation

### 3.2: Least Squares Regression Line

---

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.

## 3.1: Scatter Diagrams and Correlation

### Bivariate Data and Scattered Plot

In this chapter, we will analyze situations in which variables  $x$  and  $y$  exhibit such a linear relationship with randomness. The level of randomness will vary from situation to situation. In the introductory example connecting an electric current and the level of carbon monoxide in air, the relationship is almost perfect. In other situations, such as the height and weights of individuals, the connection between the two variables involves a high degree of randomness. In the next section, we will see how to quantify the strength of the linear relationship between two variables.

Bivariate data consists of **paired values** for two variables measured on the same individuals or objects. The goal is usually to study the relationship between the two variables.

#### Explanatory Variable and Response Variables

Bivariate data consists of two variables. An explanatory variable is the variable used to explain or predict changes in another variable. It is also called the independent (input) variable and is usually denoted by  $x$ .

The response variable is the variable that measures the outcome or response to changes in the explanatory variable. It is also called the dependent (output) variable and is usually denoted by  $y$ .

Suppose we study the relationship between hours studied and exam scores:

- **Explanatory variable ( $x$ ):** Hours studied
- **Response variable ( $y$ ):** Exam score

The exam score changes in response to the number of hours studied.

The following table provides examples of variable pairs that could be of interest from a statistical perspective.

Table 3.1.1: Explanatory Variable and Response Variable

|                   | Predictor or independent variable<br>( Explanatory Variable) $x$ | Response or dependent variable<br>(Response Variable) $y$ |
|-------------------|------------------------------------------------------------------|-----------------------------------------------------------|
| <b>Scenario 1</b> | Height of the children                                           | Head circumference                                        |
| <b>Scenario 2</b> | Area of a house (sq.ft.)                                         | Value of the house                                        |
| <b>Scenario 3</b> | Age of a particular make and model car                           | Resale value of the car                                   |
| <b>Scenario 4</b> | Amount spent by a business on advertising in a year              | Revenue received that year                                |
| <b>Scenario 5</b> | Height of a 25-year-old man                                      | The weight of the man                                     |

Consider the following table with explanatory variable  $x$  and reponse variable  $y$ .

|     |     |     |    |    |     |
|-----|-----|-----|----|----|-----|
| $x$ | -40 | -15 | 0  | 20 | 50  |
| $y$ | -40 | 5   | 32 | 68 | 122 |

We can plot these data by choosing a pair of perpendicular lines in the plane, called the coordinate axes, as shown in Figure 3.1.1 . Then to each pair of numbers in the table we associate a unique point in the plane, the point that lies  $x$  units to the right of the vertical axis (to the left if  $x < 0$ ) and  $y$  units above the horizontal axis (below if  $y < 0$ ). The relationship between  $x$  and  $y$  is called a linear relationship because the points so plotted all lie on a single straight line. The diagram is called a scatter plot.

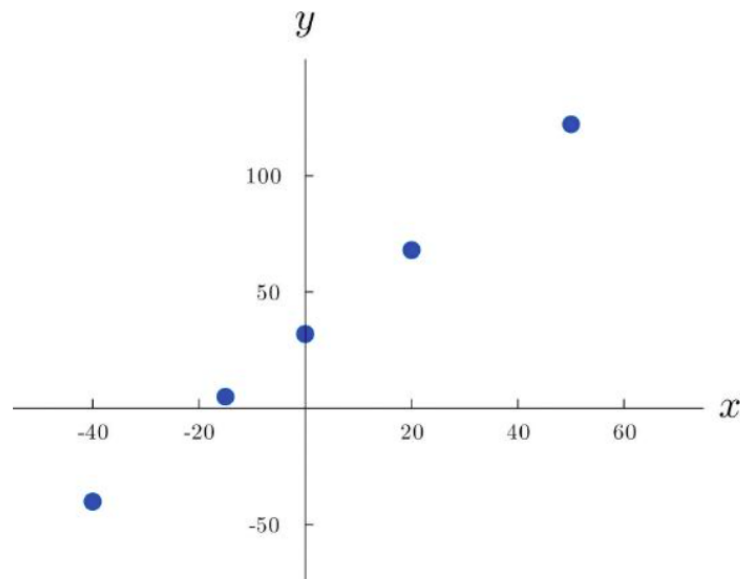


Figure 3.1.1: Plot of Celsius and Fahrenheit Temperature Pairs

## Scattered Diagram

A **scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram.

The relationship between  $x$  and  $y$  in the temperature example is deterministic because once the value of  $x$  is known, the value of  $y$  is completely determined. In contrast, all the other relationships listed in the table above have an element of randomness in them.

## Scattered Diagram Examples

### Example 1

Consider the relationship described in the last line of the table, the height  $x$  of a man aged 25 and his weight  $y$ . If we were to randomly select several 25-year-old men and measure the height and weight of each one, we might obtain a collection of  $(x, y)$  pairs something like this:

|              |     |     |     |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $x$ (height) | 68  | 72  | 69  | 72  | 70  | 73  | 70  | 73  | 71  | 74  | 72  | 75  |
| $y$ (weight) | 151 | 163 | 146 | 180 | 157 | 170 | 164 | 175 | 171 | 178 | 160 | 188 |

A scatter plot of these data is shown in Figure 3.1.3.

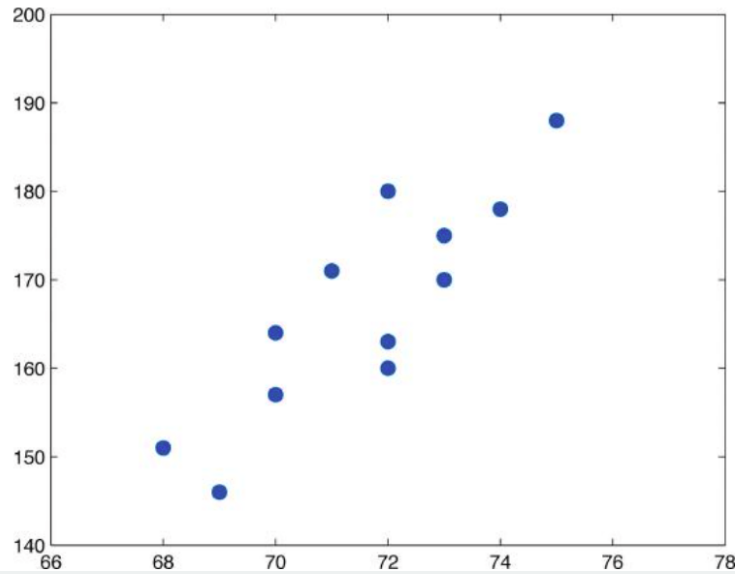


Figure 3.1.3: Scattered Diagram Study Hours vs Exam Score.

Looking at the plot, it is evident that there exists a linear relationship between height  $x$  and weight  $y$ , but not a perfect one as we see above. The points appear to be following a line, but not exactly. There is an element of randomness present.

## Example 2

A statistics teacher recorded weekly study hours ( $x$ ) and final exam scores ( $y$ ) for a group of 8 students:

Table 3.1.2: Study Hours Vs Final Exam Score

| Student name | Weekly Study Hours | Final Exam score |
|--------------|--------------------|------------------|
| A            | 2                  | 58               |
| B            | 3                  | 62               |
| C            | 4                  | 68               |
| D            | 5                  | 74               |
| E            | 6                  | 72               |
| F            | 7                  | 79               |
| G            | 8                  | 85               |
| H            | 9                  | 88               |
| I            | 10                 | 90               |
| J            | 11                 | 93               |
| K            | 12                 | 96               |

The scatter plot of the data table 3.1.2 above is shown in the following diagram.

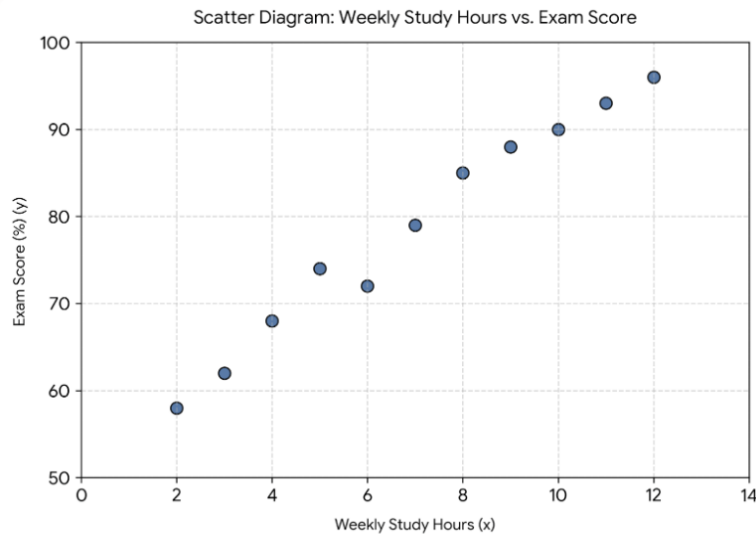


Figure 3.1.3: Scattered Diagram Study Hours vs Exam Score

### Example 3

The dataset with a negative association, using the example of car value versus age. As the graph shows, as the age of the vehicle increases, its resale value typically decreases.

Table Table 3.1.3: Age of Car Vs Resale Value

| Car | Age (X, Years) | Resale Value (Y, USD) |
|-----|----------------|-----------------------|
| A   | 2              | \$33,000              |
| B   | 3              | \$31,000              |
| C   | 4              | \$27,000              |
| D   | 6              | \$25,000              |
| E   | 8              | \$22,000              |
| F   | 10             | \$18,800              |
| G   | 12             | \$16,000              |
| H   | 14             | \$12,800              |
| I   | 16             | \$8,800               |
| J   | 18             | \$5,000               |

The scatter plot of the data table 3.1.3 above is shown in the following diagram.

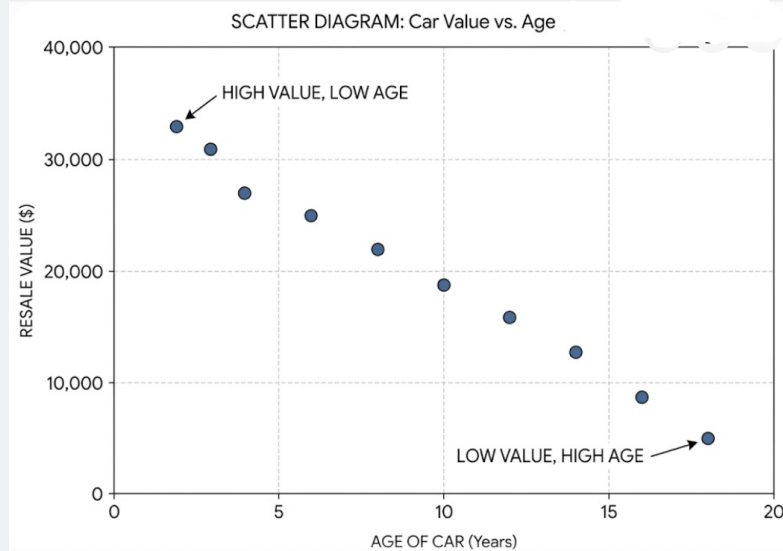


Figure 3.1.4: Scattered Diagram Age of Car vs Resale Value

### Your Turn 3.1.1: Draw Scattered Plot

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Relationship Between Two Variables

### Positively Associated and Negatively Associated

Positively associated: If the value of one variable increases, the value of the other variable also increases.

Negatively associated: If the value of one variable increases, the value of the other variable decreases.

Figure 3.1.5 illustrates linear relationships between two variables  $x$  and  $y$  of varying strengths. It is visually apparent that in the situation in panel (a),  $x$  could serve as a useful predictor of  $y$ , it would be less useful in the situation illustrated in panel (b), and in the situation of panel (c), the linear relationship is so weak as to be practically nonexistent. The *linear correlation coefficient* is a number computed directly from the data that measures the strength of the linear relationship between the two variables  $x$  and  $y$ .

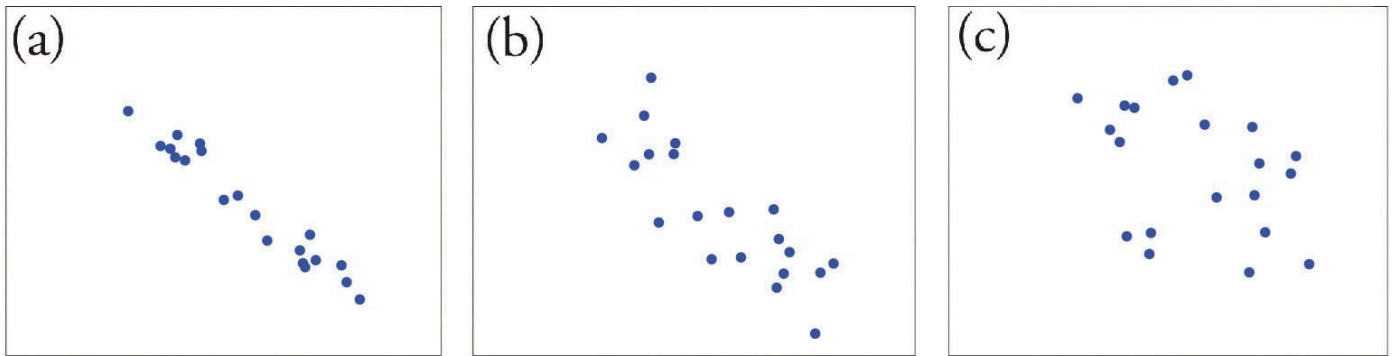


Figure 3.1.5: Linear Relationships of Varying Strengths

### Definition: Linear Correlation Coefficient

The *linear correlation coefficient* for a collection of  $n$  pairs  $x$  of numbers in a sample is the number  $r$  given by the formula

The linear correlation coefficient has the following properties, illustrated in table 3.1.6.

1. The value of  $r$  lies between  $-1$  and  $1$ , inclusive.
2. The sign of  $r$  indicates the direction of the linear relationship between  $x$  and  $y$ .
3. The value of  $|r|$  indicates the strength of the linear relationship between  $x$  and  $y$ .

It can be useful to interpret and describe the strength of a correlation. Though there are no definite rules on how the strength of a specific value must be described, there are general guidelines that can be used. These provide approximate ranges and descriptions of their strengths, but we must keep in mind that they are guidelines, and that the most appropriate wording can depend on the implications of the correlation or other considerations in research and theory relative to the variables. Thus, the cutoffs and terms used here should be considered useful but not obligatory or absolute in their meaning.

Table 3.1.6: Guidelines to describe the strength of a correlation

|    | Value of $r$         | Interpretation                            | Value of $r$       |
|----|----------------------|-------------------------------------------|--------------------|
| 1. | $r = 1$              | Perfect Positive Correlation              | $r = -1$           |
| 2. | $0.80 \leq r$        | Very Strong Positive Correlation          | $r \leq -0.80$     |
| 3. | $0.60 \leq r < 0.80$ | Strong Positive Correlation               | $\backslash(-0.80$ |
| 4. | $0.40 \leq r < 0.60$ | Moderate Positive Correlation             | $\backslash(-0.60$ |
| 5. | $0.20 \leq r < 0.40$ | Weak Positive                             | $\backslash(-0.40$ |
| 6. | $0 \leq r < 0.20$    | No Correlation to Very Weak or negligible | $\backslash(-0.20$ |

The number quantifies what is visually apparent from Figure 3.1.6. Weights tend to increase linearly with height ( $r$  is positive), and although the relationship is not perfect, it is reasonably strong ( $r$  is near 1).

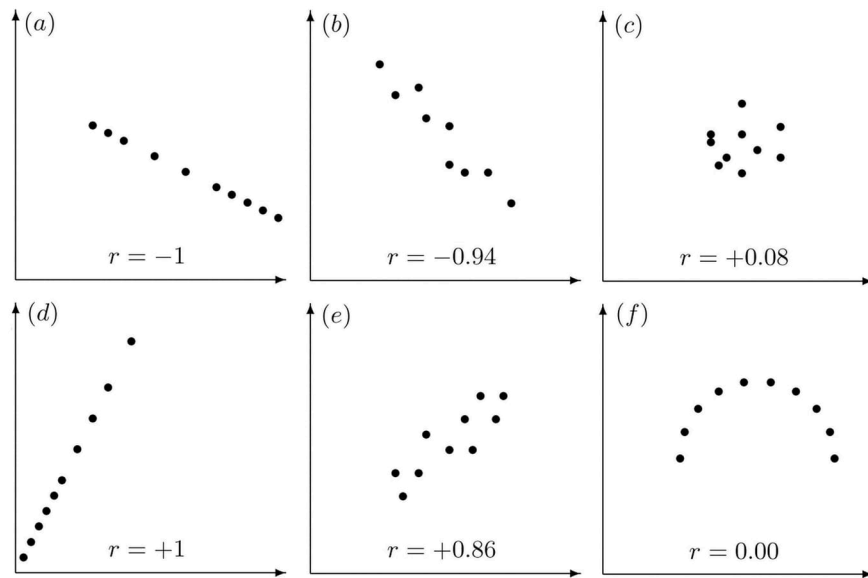


Figure 3.1.6: Linear Correlation Coefficient  $r$

Pay particular attention to panel (f), which shows a perfectly deterministic relationship between  $x$  and  $y$ , but  $r = 0$  because the relationship is not linear. (In this particular case, the points lie on the top half of a circle.)

### Your Turn 3.1.2: Strength and Direction of the Relationship

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Guess the Correlation Coefficient

## Calculating Linear Correlation Coefficient

There are several formulas for computing Pearson's correlation. Some formulas make more conceptual sense, whereas others are easier to actually compute. We will begin with a formula that makes more conceptual sense.

### Critical Value and Linear Correlation Coefficient

Table 3.1.7: Purpose or  $r$ . Why do we need to calculate it?

| Concept                      | Purpose                                                                        |
|------------------------------|--------------------------------------------------------------------------------|
| Correlation coefficient, $r$ | Measures the strength and direction of strength                                |
| Critical value, $CV$         | Decides if there is a linear relationship and depends on the sample size $n$ . |

### How to Check If There is a Linear Relationship Between Two Variables or not?

You compare the **absolute value** of the correlation coefficient to the critical value.

- If  $|r| > CV$ : There is a linear relationship between the two variables.
- If  $\leq |r|$

Table 3.1.6: Critical Value Table

| $n$  | 3     | 4    | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13    | 14    | 15    |
|------|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $CV$ | 0.997 | 0.95 | 0.878 | 0.811 | 0.754 | 0.707 | 0.666 | 0.632 | 0.602 | 0.576 | 0.553 | 0.532 | 0.514 |

### Formula to Find Linear Correlation Coefficient

We use the following formula to calculate the linear correlation coefficient,  $r$ .

$$r = \frac{\sum XY}{\sqrt{\sum x^2 \sum Y^2}}$$

- $n$ : Total number of pairs
- $X = x - \bar{x}$ : Deviation of  $x$
- $Y = y - \bar{y}$ : Deviation of  $y$
- $\sum x$ : Sum of all  $x$  values
- $\sum y$ : Sum of all  $y$  values
- $\sum X^2$ : Sum of all the squares of the  $x$  values
- $\sum Y^2$ : Sum of all the squares of the  $y$  values
- $\sum XY$ : Sum of all the products of the corresponding  $x$  and  $y$  values

But you can use the attached Online Statistics calculator to find  $r$ . Please watch the attached video. I also provided the steps in Example 3.1.3.

### Example 3.1.3: Calculate $r$

Find the

Table 3.1.7: Data Table

|     |   |   |    |    |    |
|-----|---|---|----|----|----|
| $x$ | 1 | 3 | 5  | 5  | 6  |
| $y$ | 4 | 6 | 10 | 12 | 13 |

Answer available at the end of this section.

### Your Turn 3.1.4: Calculate $r$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Coefficient of Determination

### Coefficient of Determination, $r^2$

Think of  $r^2$  as answering one practical question:

**“How much of what I’m seeing can my model actually explain?”**

Imagine you're trying to predict house prices based only on **size (square footage)**.

You calculate  $r^2 = 0.75$

About 75% of the differences in house prices can be explained by house size.

The remaining 25% is due to other factors, like: location, condition, nearby schools, market trends, etc.

So size matters a lot—but it's not the whole story.

Suppose you predict weight loss based on the number of hours of exercise per week.

$r^2 = 0.60$ , it means:

About 60% of the weight loss differences are explained by exercise. The remaining 40% is due to diet, sleep, genetics, stress, etc.

### Your Turn 3.1.5: Calculate $r$ and $r^2$

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Calculator Collection

[GeoGebra Statistics Calculator](#)

## Statistic Calculator

## Video For Section 3.1

### YouTube Video

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



## YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



## Answers

### Example 3.1.3: Calculate $r$

Pearson's  $r$  is designed so that the correlation between height and weight is the same whether height is measured in inches or in feet.

Table 3.1.8: In order to calculate  $r$

| $x$           | $x$           | $X = x - \bar{x}$ | $Y = y - \bar{y}$ | $XY$           | $X^2$           | $Y^2$           |
|---------------|---------------|-------------------|-------------------|----------------|-----------------|-----------------|
| $\sum x = 20$ | $\sum y = 45$ | 0                 | 0                 | $\sum XY = 30$ | $\sum X^2 = 16$ | $\sum Y^2 = 60$ |
| 5             | 10            | 1                 | 1                 | 1              | 1               | 1               |
| 1             | 4             | -3                | -5                | 15             | 9               | 25              |
| 3             | 6             | -1                | -3                | 3              | 1               | 9               |
| 5             | 12            | 1                 | 3                 | 3              | 1               | 9               |
| $\bar{x} = 4$ | $\bar{y} = 9$ | 0                 | 0                 | 6              |                 |                 |
| 6             | 13            | 2                 | 4                 | 8              | 4               | 16              |

To achieve this property, Pearson's correlation is computed by dividing the sum of the  $XY$  column ( $\sum XY$ ) by the square root of the product of the sum of the  $X^2$  column ( $\sum X^2$ ) and the sum of the  $Y^2$  column ( $\sum Y^2$ ). The resulting formula is:

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

Therefore

$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968$$

In the online Statistics Calculator, do the following steps

1. Click List Editor
  2. Put  $x$  data in L1 and  $y$  data in L2 in order
  3. Click Clac and click Linear Regression
  4. Type Lin\_reg(L1,L2) and click calculate.
- You will find the value of  $r$  and  $r^2$ .

This page titled [3.1: Scatter Diagrams and Correlation](#) is shared under a [CC BY-NC-SA 3.0](#) license and was authored, remixed, and/or curated by Anonymous via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [Template:TranscludeAutoNum](#) has no license indicated.

## 3.2: Least Squares Regression Line

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = b + ax$$

where  $a$  and  $b$  are constant numbers. The variable  $x$  is the *independent variable or explanatory variable*, and  $y$  is the *dependent variable (response variable)*. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

The graph of a linear equation of the form  $y = b + ax$  is a **straight line**. Any line that is not vertical can be described by this equation.

For the linear equation  $y = b + ax$ ,  $a$  = slope and  $b$  =  $y$ -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  $(0, b)$  where the line crosses the  $y$ -axis.

### What is the Slope and $y$ -intercept of a Line?

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. Recall the slope formula:

$$\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{\text{change in } y}{\text{change in } x}$$

The **slope** tells us how the response variable ( $y$ ) changes for every one-unit increase in the explanatory variable ( $x$ ), on average.

Recall the  $y$ -intercept:

The  $y$ -intercept is used to describe the response variable when the explanatory variable equals zero. In other words,  $y$ -intercept is the  $y$  value when  $x = 0$ .

### Common Example of Linear Model

Table 3.2.1: Common Examples of Linear Model

| Relationship                                            | Vairables                                                |           |
|---------------------------------------------------------|----------------------------------------------------------|-----------|
| <b>Distance from a Wi-Fi router and signal strength</b> | Explanatory: distance (feet)<br>Response: Strength (dBm) | Increase  |
| <b>Outside temperature and heating bill</b>             | Explanatory: Temperature (F)<br>Response: Cost (dollar)  | For every |

| Relationship                               | Vairables                                                  |         |
|--------------------------------------------|------------------------------------------------------------|---------|
| <b>Hours of study and test scores</b>      | Explanatory: time (hours)<br>Response: test score (points) | 10 poin |
| <b>Value of the car and age of the car</b> | Explanatory: time (years)<br>Response: value (dollar)      | A car d |

### Example 3.2.1: Interpret the Slope and $y$ -intercept

A researcher wishes to examine the relationship between years of college and yearly income. Her research discovers a linear relationship  $y = 2.6x + 36$ , where  $x$  is the number of years of college completed and  $y$  is the yearly income in thousands of dollars. Interpret the slope of the line and its  $y$ -intercept.

*Answer available at the end of this section.*

### Your Turn 3.2.1: Interpret the Slope and $y$ -intercept

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## The Regression Line

Given any collection of pairs of numbers (except when all the  $x$ -values are the same) and the corresponding scatter diagram, there always exists exactly one straight line that fits the data better than any other, in the sense of minimizing the sum of the squared errors. It is called the least squares regression line. Moreover, there are formulas for its slope and  $y$ -intercept.

### Equation of the Regression Line (Prediction Line)

The equation of regression line is on the format

$$\hat{y} = b + ax$$

$\hat{y}$ : Read as  $a$  hat

Where  $a$  is the slope of the line and  $b$  is the  $y$ -intercept of the line.

Note: The least-squares regression line is also called the **line of best fit**, the **regression line**, or the **prediction line**.

## Observed Value and Predicted Value

### Observed Value $y$ :

The observed value is the actual data point you collected. It represents the real-world outcome recorded in your dataset for a specific input (explanatory variable). Observed value is usually denoted as  $y$ . In a scatter plot, the observed value is the actual "dot" or point.

### Predicted Value $\hat{y}$ :

The predicted value (often called " $y$ -hat") is the value that your regression model calculates for a given  $x$ . We calculated the observed value using the regression equation:  $\hat{y} = ax + b$ . On a scatter plot, the predicted value is the point directly on the regression line for a specific  $x$ .

## Residual and Interpretation

In regression analysis, a **residual** is the difference between an observed value and the value predicted by a mathematical model. It essentially measures the "error" for a single data point.

$$\text{Residual} = y - \hat{y}$$

$y$ : The actual observed value (the data point).

$\hat{y}$ : The predicted value (the point on the regression line).

Interpretation of Residual:

**Positive Residual:** The actual data point is above the prediction line. So that observed value is above the average.

**Negative Residual:** The actual data point is below the prediction line. So that observed value is below the average.

**Zero Residual:** The actual data point is on the prediction line. So that observed value is the average.

## Example 3.2.2: Predicted Value and Residual

Alex is investigating how long her phone's battery lasts (in hours) for various brightness levels (on a scale of 1 to 100). Alex collected data, which is displayed in the table. Do not round between steps. Round answers to two decimal places.

Table 3.2.2: for Brightness Level ( $x$ ), Hours ( $y$ )

| Brightness Level $x$ | 27  | 34  | 50  | 66  | 66  | 68  | 70  | 88  | 94  | 96  | 97  | 99  |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Hours $y$            | 6.1 | 6.1 | 5.5 | 3.3 | 4.8 | 3.8 | 5.1 | 2.9 | 2.3 | 1.9 | 2.6 | 1.9 |

The equation of the best-fit line is  $\hat{y} = -0.0608x + 8.1884$ . Use the model to answer the questions below.

- Find the observed  $y$ -value when  $x = 70$ .
- Find the predicted  $y$ -value when  $x = 70$ .
- Find the residual for this point  $x = 70$  and interpret it.

Answer available at the end of this section.

### Your Turn 3.2.2: Predicated Value and Residual

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### How to Determine the Equation of the Least Squares Regression Line?

Given a collection of pairs  $(x, y)$  of numbers (in which not all the  $x$ -values are the same), there is a line  $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$  that best fits the data in the sense of minimizing the sum of the squared errors. It is called the *least squares regression line*. Its slope  $\hat{\beta}_1$  and  $y$ -intercept  $\hat{\beta}_0$  are computed using the formulas

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$SS_{xx} = \sum x^2 - \frac{1}{n} \left( \sum x \right)^2$$

$$SS_{xy} = \sum xy - \frac{1}{n} \left( \sum x \right) \left( \sum y \right)$$

$\bar{x}$  is the mean of all the  $x$ -values,  $\bar{y}$  is the mean of all the  $y$ -values, and  $n$  is the number of pairs in the data set.

The equation

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

The equation of the least-squares regression line is called the least-squares regression equation.

**Note:** We can find the equation of the least-squares regression line using an online statistics calculator.

Please watch the attached video at the bottom of the page.

### Example 3.2.3: Find the least Square Regression Line

Table 3.2.3 shows the age in years and the retail value in thousands of dollars of a random sample of ten automobiles of the same make and model.

1. Construct the scatter diagram.
2. Compute the linear correlation coefficient  $r$ . Interpret its value in the context of the problem.
3. Compute the least squares regression line. Plot it on the scatter diagram.
4. Interpret the meaning of the slope of the least squares regression line in the context of the problem.
5. Suppose a four-year-old automobile of this make and model is selected at random. Use the regression equation to predict its retail value.

6. Calculate the residual for  $x = 4$ . Is the observed value for this  $x$  value above or below the average?
7. Suppose a 20-year-old automobile of this make and model is selected at random. Is it reasonable to predict using the retail value using the regression equation? Why or why not? Interpret the result.
8. Comment on the validity of using the regression equation to predict the price of a brand new automobile of this make and model.

Table 3.2.3: Data on Age and Value of Used Automobiles of a Specific Make and Model

| $x$ | 2    | 3    | 3    | 3    | 4    | 4    | 5    | 5    | 5    | 6    |
|-----|------|------|------|------|------|------|------|------|------|------|
| $y$ | 28.7 | 24.8 | 26.0 | 30.5 | 23.8 | 24.6 | 23.8 | 20.4 | 21.6 | 22.1 |

Answer 1 available at the end of this section.

### Your Turn 3.2.3: Compute the Least Square Regression Line

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Definition: Extrapolation and Interpolation

The process of using the least squares regression equation to estimate the value of  $y$  at a value of  $x$  that does not lie in the range of the  $x$ -values in the data set that was used to form the regression line is called extrapolation.

It is an invalid use of the regression equation that can lead to errors, hence it should be avoided.

While it is a common tool in forecasting and modeling, it carries a higher risk than **interpolation** (predicting values *within* the existing data range) because it assumes that established patterns will continue indefinitely into unknown territory.

**Interpolation:** estimating values *within the range* of the data. or within the range of observed values.

**Extrapolation:** projecting beyond the measured data or observed data (with caution!). Sometimes, extrapolation is also referred to as “**Outside the scope of the model**”.

In example 3.2.3, to make a prediction for  $x = 4$  is interpolation and making a prediction for  $x = 20$  is extrapolation

### Your Turn 3.2.4: Extrapolation and Interpolation

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Videos for Section 3.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



## Answers

### Example 3.2.1: Interpret the Slope and $y$ -intercept

The explanatory variable ( $x$ ) is the number of years of college completed, and the response variable ( $y$ ) is yearly income in thousands of dollars.

$$\text{slope} = 2.6 = \frac{2.6}{1} = \frac{\text{change in income}}{\text{change in number of years}}$$

For each additional year of college, income increases by an average of 2.6 thousand (\$2,600) dollars.

The  $y$ -intercept is ( $b = 36$ ). In other words, (0, 36) is the  $y$ -intercept.

Or for addition years of college completed, the yearly income is expected to increase by (\$2,600) dollars.

That is, if the person did not attend college (i.e., 0 years of college), his yearly income is expected to be \$36,000.

### Example 3.2.2: Predicted Value and Residual

a. Observed  $y$ -value is given in the table with is 5.1 for  $x = 70$ , so  $y = 5.1$ .

b. Predict value  $\hat{y}$ , we have to use equation to find it. Put  $x = 70$  into the equation

$$\begin{aligned}\hat{y} &= -0.0608x + 8.1884 \\ &= -0.0608(70) + 8.1884 \\ &= 3.9324\end{aligned}$$

c. To find the residual, we subtract the predicted value from the observed value

$$\text{Residual} = y - \hat{y} = 5.1 - 3.9324 = 1.1676$$

That means the observed value is above the average, since the residual is positive.

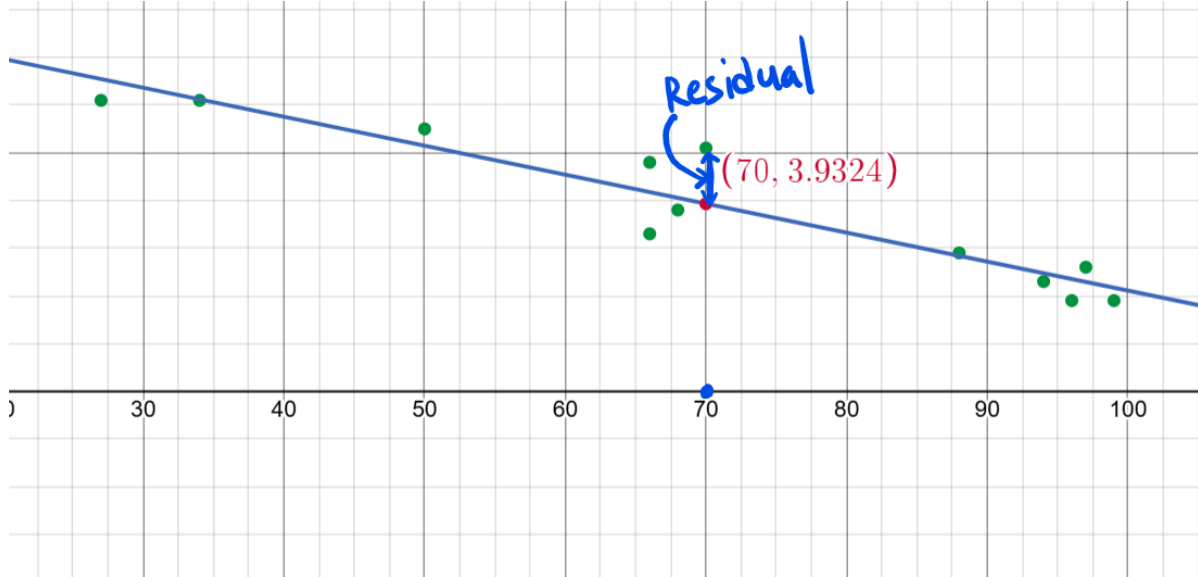


Figure 3.2.1: Residual for  $x = 70$

### Example 3.2.3: Find the least Square Regression Line

The scatter diagram is shown in Figure 3.2.2.

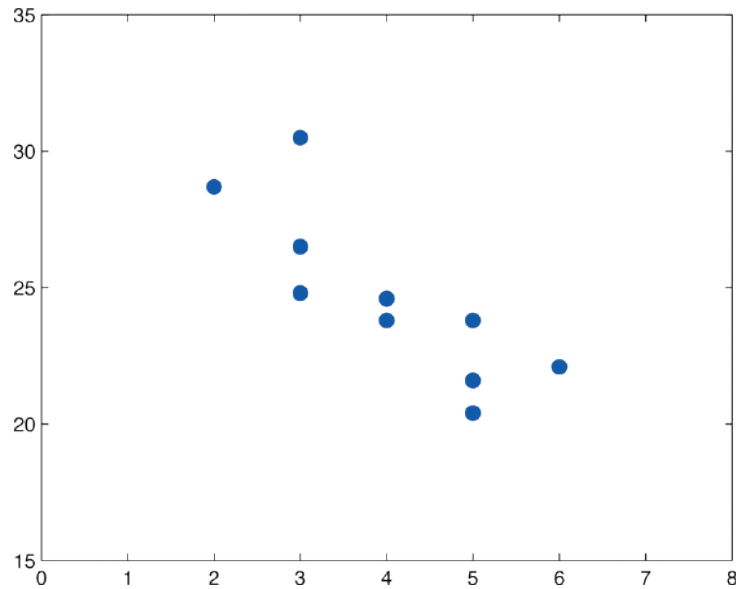


Figure 3.2.2: Scatter Diagram for Age and Value of Used Automobiles

We must first compute  $SS_{xx}$ ,  $SS_{xy}$ ,  $SS_{yy}$ , which means computing  $\sum x$ ,  $\sum y$ ,  $\sum x^2$ ,  $\sum y^2$  and  $\sum xy$ . Using a computing device we obtain

$$\sum x = 40 \quad \sum y = 246.3 \quad \sum x^2 = 174 \quad \sum y^2 = 6154.15 \quad \sum xy = 956.5$$

Thus

$$\begin{aligned}
 SS_{xx} &= \sum x^2 - \frac{1}{n} \left( \sum x \right)^2 \\
 &= 174 - \frac{1}{10} (40)^2 = 14 \\
 SS_{xy} &= \sum xy - \frac{1}{n} \left( \sum x \right) \left( \sum y \right) \\
 &= 956.5 - \frac{1}{10} (40)(246.3) = -28.7 \\
 SS_{yy} &= \sum y^2 - \frac{1}{n} \left( \sum y \right)^2 \\
 &= 6154.15 - \frac{1}{10} (246.3)^2 = 87.781
 \end{aligned}$$

so that

$$\begin{aligned}
 r &= \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} \\
 &= \frac{-28.7}{\sqrt{(14)(87.781)}} \\
 &= -0.819
 \end{aligned}$$

The age and value of this make and model automobile are moderately strongly negatively correlated. As the age increases, the value of the automobile tends to decrease.

Using the values of  $\sum x$  and  $\sum y$  computed in part (b),

$$\begin{aligned}
 \bar{x} &= \frac{\sum x}{n} = \frac{40}{10} = 4 \\
 \bar{y} &= \frac{\sum y}{n} = \frac{246.3}{10} = 24.63
 \end{aligned}$$

Thus using the values of  $SS_{xx}$  and  $SS_{xy}$  from part (b),

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-28.7}{14} = -2.05$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 24.63 - (-2.05)(4) = 32.83$$

The equation  $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$  of the least squares regression line for these sample data is

$$\hat{y} = -2.05x + 32.83$$

We can also find the least squares regression line using an online statistics calculator. Please do the following steps

In the online Statistics Calculator, do the following steps

1. Click List Editor
2. Put  $x$  data in L1 and  $y$  data in L2 in order
3. Click Clac and click Linear Regression
4. Type Lin\_reg(L1,L2) and click calculate.

You will find  $y = ax + b$  with the values of  $a$  and  $b$  gives the equation.

The slope  $-2.05$ .

$$\text{slope} = -2.05 = \frac{-2.05}{1} = \frac{\text{change in price}}{\text{change in year}}$$

**Interpretation:** For each unit increase in  $x$  (additional year of age), the average value of this make and model vehicle decreases by about 2.05 thousand dollars (about \$2,050).

Since we know nothing about the automobile other than its age, we assume that it is of about average value and use the average value of all four-year-old vehicles of this make and model as our estimate. The average value is simply the value of  $\hat{y}$  obtained when the number 4 is inserted for  $x$  in the least squares regression equation:

$$\hat{y} = -2.05(4) + 32.83 = 24.63$$

Which corresponds to  $\$24.63 \times 1000 = \$24,630$

For  $x = 4$ ,  $y = 23.8$  from table. From part (5),  $\hat{y} = 24.63$

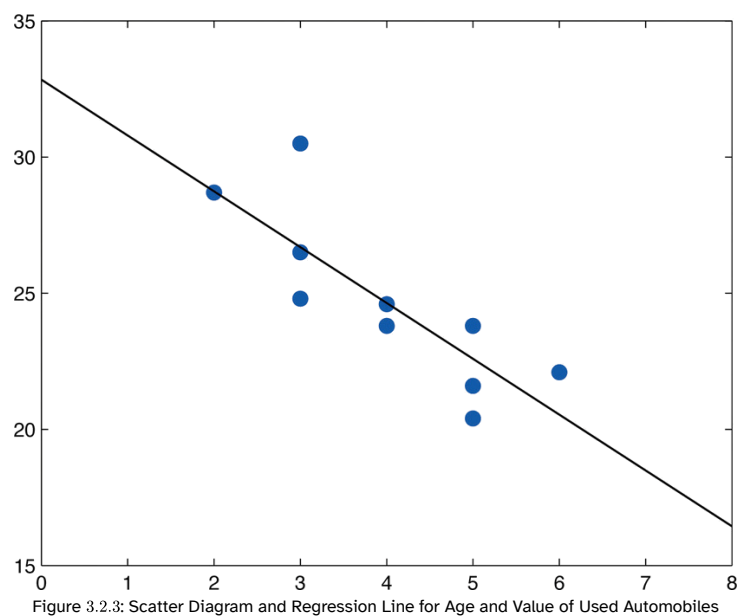
$$\text{Residual} = y - \hat{y} = 23.8 - 24.63 = -0.83$$

That means the observed value is below the average.

Now we insert  $x = 20$  into the least squares regression equation, to obtain

$$\hat{y} = -2.05(20) + 32.83 = -8.17$$

which corresponds to  $-\$8,170$ . Something is wrong here, since a negative makes no sense. The error arose from applying the regression equation to a value of  $x$  outside the range of  $x$ -values in the original data, from 2 to 6 years. It is an invalid use of the regression equation and should be avoided.



This page titled [3.2: Least Squares Regression Line](#) is shared under a license and was authored, remixed, and/or curated by Anonymous via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [11: Sequences and Series](#) by David Guichard is licensed [CC BY-NC-SA 4.0](#). Original source: <https://www.whitman.edu/mathematics/multivariable/>.

## 4: PROBABILITY

- 4.1: Fundamental Concepts of Probability
- 4.2: The Addition and Complement Rules
- 4.3: Independence and Multiplication Rules

---

This page titled [4: Probability](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Kathryn Kozak](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.5: Use the Properties of Logarithms](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.

## 4.1: Fundamental Concepts of Probability

---

### Introducing Probability

It all comes down to this. The game *Monopoly*, which started hours ago, is in the home stretch. Your sister has the dice, and if she rolls a 4, 5, or 7, she'll land on one of your best spaces, and the game will be over. How likely is it that the game will end on the next turn? Is it more likely than not? How can we measure that likelihood? This section addresses this question by introducing a way to measure uncertainty.

If you roll a die, pick a card from the deck of playing cards, or randomly select a person and observe their hair color, we are executing an experiment or procedure. In probability, we look at the likelihood of different outcomes. We begin with some terminology

Uncertainty is, almost by definition, a nebulous concept. In order to put enough constraints on it that we can mathematically study it, we will focus on uncertainty strictly in the context of experiments. Recall that experiments are processes whose outcomes are unknown; the sample space for the experiment is the collection of all those possible outcomes. When we want to talk about the likelihood of particular outcomes, we sometimes group outcomes together; for example, in the *Monopoly* example at the beginning of this section, we were interested in the roll of 4 dice that might fall as a 4, 5, or 7. A grouping of outcomes that we're interested in is called an event. In other words, an event is a subset of the sample space of an experiment; it often consists of the outcomes of interest to the experimenter.

When we are talking about combinatorics or probability, the word "experiment" has a slightly different meaning than it does in the sciences. Experiments can range from very simple ("flip a coin") to very complex ("count the number of uranium atoms that undergo nuclear fission in a sample of a given size over the course of an hour"). Experiments have unknown outcomes that generally rely on chance, so if the experiment is repeated (or replicated), the outcome might be different. No matter the experiment, analysis typically begins by identifying its sample space.

#### Probability

Probability is the measure of the chance of an event occurring. Probability takes a numerical value that represents how likely an event is to occur. It is always a value between 0 (impossible) and 1 (certain).

#### Experiment

In probability, an experiment is any process or action that can be repeated, but where the exact outcome isn't known in advance.

Example of a probability experiment:

1. Tossing a coin.
2. Roll a die
3. Pick one card from a standard deck of 52.
4. Testing a lightbulb from a factory line to see if it is "defective" or "non-defective."

5. Observing whether it rains or stays dry on a specific day.

### Outcome

The possible results of an experiment are called **outcomes**. For example, in the experiment of tossing a coin, possible outcomes are head and tail.

### Sample Space

The **sample space** of an experiment is the set of all possible outcomes of the experiment.

It's often expressed as a set (i.e., as a list bound by braces)

If the experiment is *a roll, a single die is rolled once*, the sample space would be written  $\{1, 2, 3, 4, 5, 6\}$ .

### Event

An **event** is any particular outcome or group of outcomes. The event is a subset of the sample space.

If the experiment is *a roll, a single die is rolled once, the event could be "roll an even number" or  $\{2, 4, 6\}$ .*

### Example 4.1.1: Finding the Sample Space (Dice and Coin)

For each of the following experiments, identify the sample space.

1. Flip a coin (which has 2 faces, typically called "heads" and "tails").
2. Roll a 6-sided die.
3. Roll two 6-sided dice.

*Answer available at the end of this section.*

### Your Turn 4.1.1: Find Sample Space

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



A standard deck of cards includes:

4 suits: Hearts, Diamonds, Clubs, Spades

13 cards in each suit. So total  $13 \times 4 = 52$  cards (excluding Jokers).

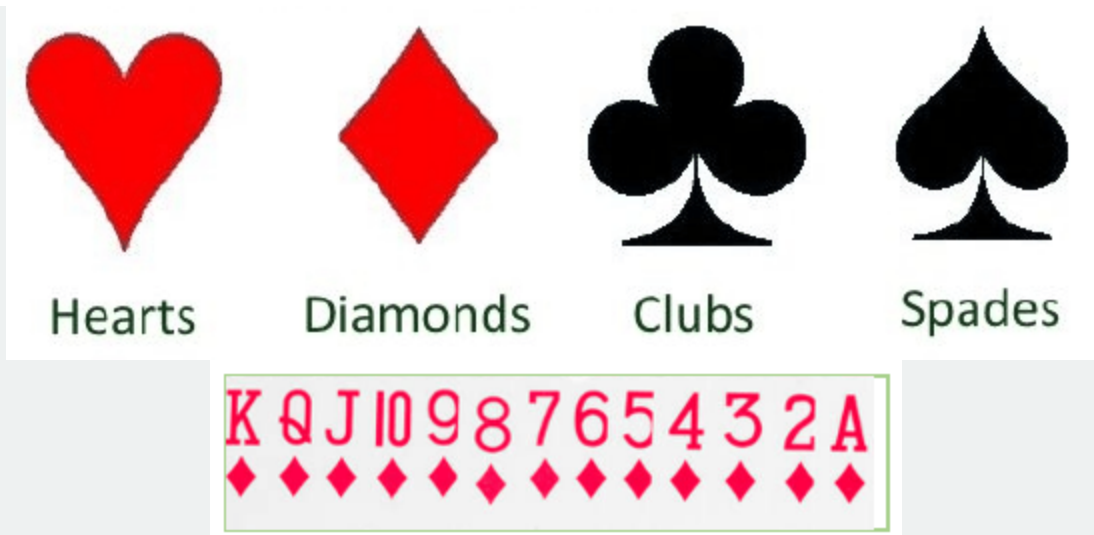


Figure 4.1.2A: Suit of Cards (4 Suit) (TOP), Rank of cards (13 rank) (BOTTOM)

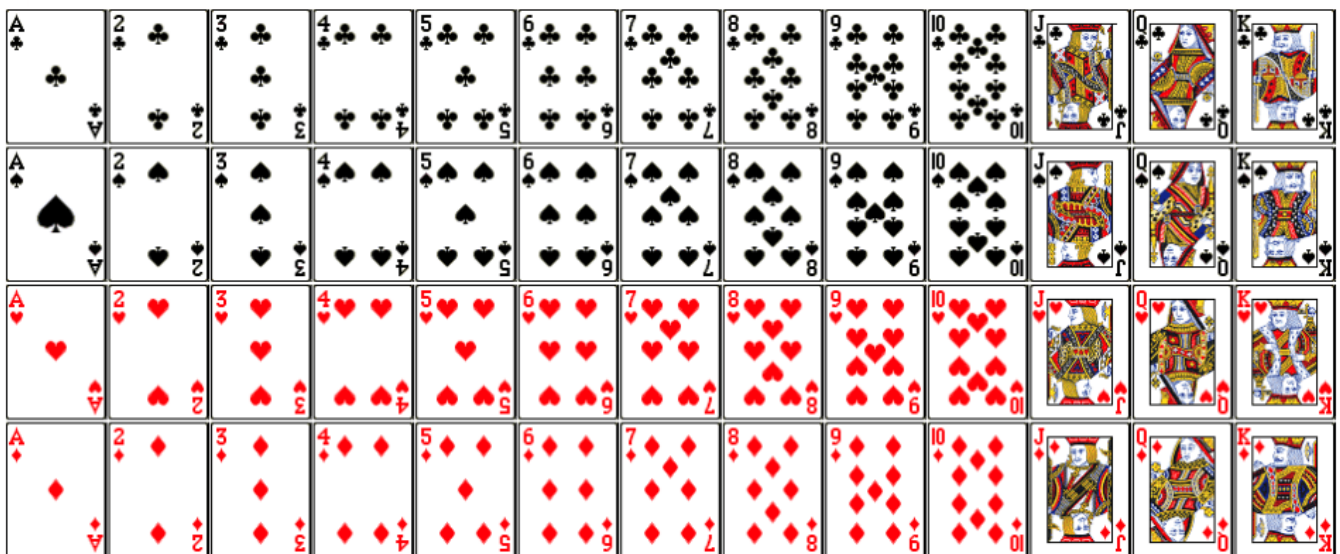


Figure 4.1.2B: Deck of 52 cards

### Example 4.1.2: Finding the Event (Die and Deck of Cards)

Identify the event in the following experiment.

1. Roll one 6-sided die. Find the event rolling an even number.
2. Roll two 6-sided dice. Find the event rolling two same numbers.
3. Draw a card from a deck of 52 cards. Find the event drawing king.

*Answer available at the end of this section.*

## Using Tree Diagrams to Identify Sample Spaces

In experiments with two or more stages, or when the stages are dependent, a tree diagram is a helpful tool for systematically identifying the sample space. Tree diagrams are built by first drawing a single point (or *node*), then, from that node, drawing one branch (a short line segment) for each outcome of the first stage. Each branch gets its own node at the other end (which we typically label with the corresponding outcome for that branch); from each of these, we draw another branch for each outcome

of the second stage, assuming that the outcome of the first stage matches the branch we were on. If there are other stages, we can continue from there by continuing to add branches and nodes. This sounds really complicated, but it's easier to understand through an example.

Example: when we flip two coins, the Sample space has four outcomes: {HH, HT, TH, TT}. Note that each sequence shows the result of the **first flip** followed by the **second flip**.

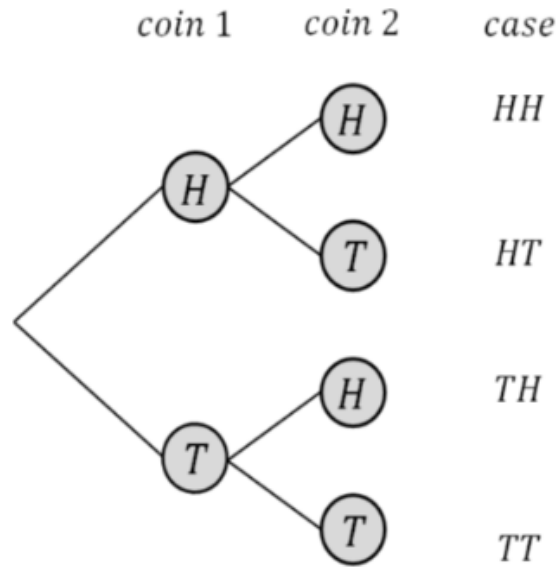


Figure 4.1.2: Flip Two Coins

### Example 4.1.3: Finding the Sample Space Using Tree Diagram

Use a tree diagram to find the sample spaces of each of the following experiments:

1. You flip a coin 3 times.
2. You flip a coin. If the result is a head, you roll a 4 -sided die. If it's a tail, you roll a 6-sided die.

*Answer available at the end of this section.*

### Your Turn 4.1.3: Identify Sample Space

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Theoretical Probability

The theoretical method gives the most reliable results, but cannot always be used. Suppose the sample space of an experiment consists of equally likely outcomes. In that case, the theoretical probability of an event is defined as the ratio of the number of

outcomes in the event to the number of outcomes in the sample space.

### Probability Notation: Possible Values of Probability

For any event  $A$ , the notation  $P(A)$  is for the probability of event  $A$  occurring. The smallest possible probability is 0, if no outcomes correspond with the event. The largest possible probability is 1, if all possible outcomes correspond with the event.

The probability of any event must be between zero and one, inclusive. That is

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

Note: You can express **probability as a percentage** by simply converting the decimal value to a percentage. For example, the probability of tossing heads when we flip one coin is 0.5. We can also say that when we flip a coin, there is a 50% chance that the coin lands as a head.

In the course of this chapter, if you compute a probability and get an answer that is negative or greater than 1, you have made a mistake and should check your work.

An **impossible event** has a probability of 0. In another word, if  $P(A) = 0$ , the event  $A$  is impossible.

A **certain event** has a probability of 1 or 100%. In another word, if  $P(A) = 1$ , the event  $A$  is certain.

Probabilities are essentially fractions and can be reduced to lower terms, like fractions, or expressed as decimals.

### Your Turn 4.1.4: Classify the Event

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### FORMULA: Theoretical Probability

For an experiment whose sample space  $S$  consists of equally likely outcomes, the **theoretical probability** of the event  $E$  is the ratio

$$P(\text{Event}) = \frac{\text{number of outcomes in the event}}{\text{number of outcomes in sample space}} = \frac{\text{How many outcomes in the event?}}{\text{How many outcomes in the sample space?}}$$

### Example 4.1.5: Compute Theoretical Probability (Die)

If we roll a 6-sided die, calculate

1.  $P(\text{rolling a 1})$ .
2.  $P(\text{rolling a number bigger than 4})$ .

Answer available at the end of this section.

### Your Turn 4.1.5: Roll a Die

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 4.1.6: Determining Certain and Impossible Events

A. Consider an experiment that consists of rolling a single standard 6-sided die (with faces numbered 1 – 6). Decide if these probabilities are equal to zero, equal to one, or somewhere in between.

1.  $P(\text{roll } 4)$ .
2.  $P(\text{roll } 7)$ .
3.  $P(\text{roll a positive number})$ .
4.  $P(\text{roll } \frac{1}{3})$ .
5.  $P(\text{roll an even number})$ .

B. Consider an experiment of drawing a card from a deck of cards. Decide if these probabilities are equal to zero, equal to one, or somewhere in between.

1. A card is a king.
2. A card is a king or queen.
3. A card is red and black.
4. A card is red or black.

Answer available at the end of this section.

### Your Turn 4.1.6: Multi Sided Die

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Your Turn 4.1.7: Deck of Cards

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 4.1.8: Spinner

Suppose the spinner shown below is used for a game. Find the following probability.



Figure 4.1.6: Spinning Wheel with Six number 1 to 6

1.  $P(\text{spin } 5)$ .
2.  $P(\text{spin multiple of } 2)$ .
3.  $P(\text{spin a number greater than } 4)$ .
4.  $P(\text{spin } 2 \text{ and } 3)$ .

*Answer available at the end of this section.*

### Your Turn 4.1.8: Spinner

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



The theoretical method gives the most reliable results, but it cannot always be used. If the sample space of an experiment consists of equally likely outcomes, then the theoretical probability of an event is defined to be the ratio of the number of outcomes in the event to the number of outcomes in the sample space.

### Checkpoint: Equally Likely for Theoretical Probability

It is critical that you make sure that every outcome in a sample space is equally likely before you compute theoretical probabilities!

### Example 4.1.9: Find Theoretical Probabilities (Roll of Two Die)

In the Basic Concepts of Probability, we were considering a *Monopoly* game where, if your sister rolled a sum of 4, 5, or 7 with 2 standard dice, you would win the game. What is the probability of this event? Use tables to determine your answer.

Answer available at the end of this section.

### Your Turn 4.1.9: Roll Two Dice

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example 4.1.10: Tree Diagrams to Compute Theoretical Probability (Flip Coins 3 times)

If you flip a fair coin 3 times, what is the probability of each event? Use a tree diagram to determine your answer

1. You flip exactly 2 heads.
2. You flip 2 consecutive heads at some point in the three flips.
3. All 3 flips show the same result.

Answer available at the end of this section.

### Your Turn 4.1.10: Roll Die Two Times

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### People in Mathematics: Gerolamo Cardano

The first known text that provided a systematic approach to probabilities was written in 1564 by Gerolamo Cardano (1501–1576). Cardano was a physician whose illegitimate birth closed many doors that would have otherwise been open to someone with a medical degree in 16th-century Italy. As a result, Cardano often turned to gambling to help ends meet. He was a remarkable mathematician, and he used his knowledge to gain an edge when playing cards or dice. His 1564 work, titled *Liber de ludo aleae* (which translates as *Book on Games of Chance*), summarized everything he knew about probability. Of course, if that book fell into the hands of those he played against, his advantage would disappear. That's why he never allowed it to be published in his lifetime (it was eventually published in 1663). Cardano made other contributions to mathematics; he was the first person to publish the third-degree analogue of the Quadratic Formula (though he didn't discover it himself), and he popularized the use of negative numbers.

## Empirical Probability

Theoretical probabilities are precise but may not be available in every situation. If the outcomes in the sample space are not equally likely, then we're out of luck. Suppose you're watching a baseball game, and your favorite player is about to step up to the plate. What is the probability that he will get a hit?

In this case, the sample space is {hit, not a hit}. That doesn't mean that the probability of a hit is  $\frac{1}{2}$ , since those outcomes aren't equally likely. The theoretical method simply can't be used in this situation. Instead, we might look at the player's statistics up to this point in the season, and see that he has 122 hits in 531 opportunities. So, we might think that the probability of a hit in

the next plate appearance would be about  $\frac{122}{531} \simeq 0.23$ . When we use the outcomes of previous replications of an experiment to assign a probability to the next replication, we're defining an empirical probability. Empirical probability is assigned using the outcomes of previous replications of an experiment by finding the ratio of the number of times the event occurred in the previous replications to the total number of previous replications.

Empirical probabilities aren't exact, but when the number of previous replications is large, we expect them to be close. Also, if the previous runs of the experiment were not conducted under the exact set of circumstances as the one we're interested in, the empirical probability is less reliable. For instance, in the case of our favorite baseball player, we might try to get a better estimate of the probability of a hit by looking only at his history against left- or right-handed pitchers (depending on the handedness of the pitcher he is about to face).

### FORMULA: EMPIRICAL PROBABILITY

Empirical probability is the probability of an event occurring based on actual experiments or historical data rather than theoretical calculations. It is calculated using the formula:

$$P(\text{Event}) = \frac{\text{number of times event occurs}}{\text{total number of trials}}$$

If you flip a coin 100 times and it lands on heads 48 times, the empirical probability of getting heads is:

$$P(\text{Heads}) = \frac{48}{100} = 0.48$$

Empirical probability differs from theoretical probability because it is based on observed outcomes rather than assumed equal likelihoods. It is useful in real-world scenarios where theoretical probabilities cannot be easily determined.

### Example 4.1.11: Finding Empirical Probabilities

Find the empirical probability of the following events:

1. Jose is on the basketball court practicing his shots from the free-throw line. He made 47 out of his last 80 attempts. What is the probability he will make his next shot?
2. Amy is about to begin her morning commute. Over her last 60 commutes, she arrived at work 12 times in under half an hour. What is the probability that she arrives at work in 30 minutes or less?
3. Felix is playing *Yahtzee* with his sister. Felix won 14 of the last 20 games he played against her. How likely is he to win this game?

*Answer available at the end of this section.*

### Your Turn 4.1.11: Empirical Probability

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 4.1.12: Empirical Probability

The number of M&Ms of each color found in a case containing 1600 M&Ms is given in the table below.

### M&M Color and its Frequency

| Color  | Freq |
|--------|------|
| blue   | 238  |
| brown  | 338  |
| green  | 419  |
| orange | 166  |
| red    | 323  |
| yellow | 116  |

- Find the percentage of yellow M&M's in the case.
- Find the probability that a randomly selected M&M will be in the category "yellow or green".

Answer available at the end of this section.

### Your Turn 4.1.12: Probability from Two Way Table

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Answers

### Example 4.1.1: Finding the Sample Space (Dice and Coin)

- If we use " $H$ " to denote "heads is facing up" and " $T$ " to denote "tails is facing up", then  
Sample space:  $\{H, T\}$
- There are 6 numbers on the die: 1, 2, 3, 4, 5, and 6, then  
Sample space:  $\{1, 2, 3, 4, 5, 6\}$
- Note that each ordered pair below shows the result of the first die followed by the second die.

|           |  | SECOND DIE |        |        |        |        |        |
|-----------|--|------------|--------|--------|--------|--------|--------|
|           |  |            |        |        |        |        |        |
| FIRST DIE |  | (1, 1)     | (1, 2) | (1, 3) | (1, 4) | (1, 5) | (1, 6) |
|           |  | (2, 1)     | (2, 2) | (2, 3) | (2, 4) | (2, 5) | (2, 6) |
|           |  | (3, 1)     | (3, 2) | (3, 3) | (3, 4) | (3, 5) | (3, 6) |
|           |  | (4, 1)     | (4, 2) | (4, 3) | (4, 4) | (4, 5) | (4, 6) |
|           |  | (5, 1)     | (5, 2) | (5, 3) | (5, 4) | (5, 5) | (5, 6) |
|           |  | (6, 1)     | (6, 2) | (6, 3) | (6, 4) | (6, 5) | (6, 6) |

Figure 4.1.1: Possible outcomes of rolling two dice

### Example 4.1.2: Finding the Event (Die and Deck of Cards)

1. The sample space is the set of all possible outcomes:  $\{1, 2, 3, 4, 5, 6\}$ . There are 6 outcomes in the sample space.

Event:  $\{2, 4, 6\}$

2. All 36 possible outcomes while rolling two dice are given in the table in the example 4.1.1.

So, for example,  $(3, 2)$  represents the outcome where one 6-sided roll results in a 3, and another 6-sided roll gives us a 2.

Event:  $\{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

3. Recall that a standard deck of cards consists of 52 unique cards which are labeled with a rank (the whole numbers from 2 to 10, plus J, Q, K, and A) and a suit ( $\clubsuit$ ,  $\diamond$ ,  $\heartsuit$ , or  $\spadesuit$ ). A standard deck is thoroughly shuffled, and you draw one card at random (so every card has an equal chance of being drawn). There are 52 possible outcomes while drawing a card from a deck of cards. (See the above figure 4.1.2C )

Event:  $\{\clubsuit K, \diamond K, \heartsuit K, \spadesuit K\}$

### Example 4.1.3: Finding the Sample Space Using Tree Diagram

Let H represent the head outcome, and T represent the tail outcome. See figure 4.1.4.

The sample space is:  $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ .

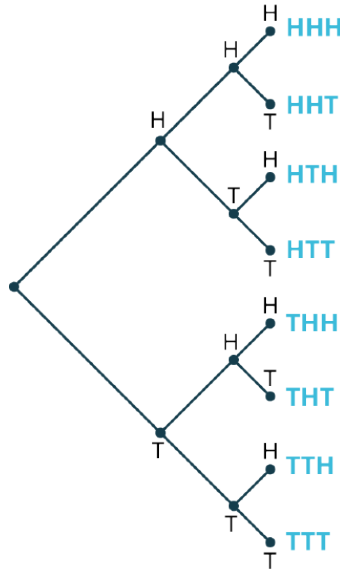


Figure 4.1.4: Sample Space of Flipping Three Coins

If the result is H, you roll either 1, 2, 3, and 4. If result is T, you roll either 1, 2, 3, 4, 5, and 6. See figure 4.1.5.

The sample space is  $\{H1, H2, H3, H4, T1, T2, T3, T4, T5, T6\}$

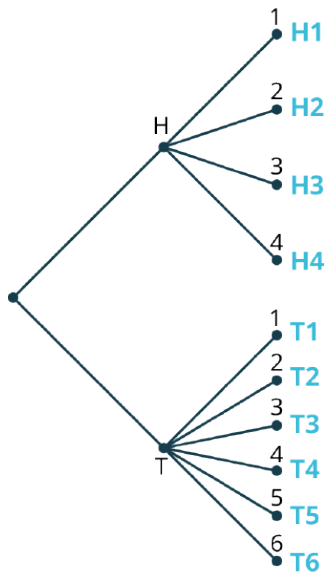


Figure 4.1.5: Sample Sapce

### Example 4.1.5: Compute Theoretical Probability (Die)

Recall that the sample space when rolling a die is  $\{1, 2, 3, 4, 5, 6\}$ .

1. There is one outcome corresponding to “rolling a 1”, so

$$P(\text{rolling a 1}) = \frac{1}{6}$$

2. There are two outcomes bigger than a 4, so

$$P(\text{rolling a number bigger than 4}) = \frac{2}{6} = \frac{1}{3}$$

#### Example 4.1.6: Determining Certain and Impossible Events

A. Let's start by identifying the sample space. For one roll of this die, the possible outcomes are  $\{1, 2, 3, 4, 5, 6\}$ . We can use that to assess these probabilities:

1. We see that 4 is in the sample space, so it's possible that it will be the outcome. It's not certain to be the outcome, though.  $P(\text{roll 4})$  is somewhere in between.
2. Notice that 7 is not in the sample space. So,  $P(\text{roll a 7}) = 0$ .
3. Every outcome in the sample space is a positive number, so this event is certain.  $P(\text{roll a positive number})$  is certain.
4. Since  $\frac{1}{3}$  is not in the sample space,  $P(\text{roll } \frac{1}{3}) = 0$ , so this is impossible event.
5. Some outcomes in the sample space are even numbers 2, 4, and 6, but the others aren't. So,  $0 < P(\text{roll an even number}) < 1$

B. There are 52 cards in the deck, so the sample space for each of these experiments has 52 elements. That will be the denominator for each of our probabilities.

1. There are 4 king cards in total in the deck, so  $P(\text{king}) = \frac{4}{52} = \frac{1}{13}$ , so this event probability is in between zero and one.
2. There are 8 king or queen cards in the deck, so  $P(\text{king or queen}) = \frac{8}{52} = \frac{2}{13}$ , so this event probability is in between zero and one.
3. There are 0 red and black cards in the deck, so  $P(\text{red and black}) = \frac{0}{52} = 0$ , so this event is impossible.
4. There are 52 red or black cards in the deck, so  $P(\text{red or black}) = \frac{52}{52} = 1$ , so this event is certain.

#### Example 4.1.8: Spinner

There are 6 possible outcomes  $\{1, 2, 3, 4, 5, 6\}$ .

1. There is only one outcome 5, so

$$P(\text{spin 5}) = \frac{1}{6}$$

2. There is only three outcomes as multiple of 2, they are 2, 4, 6, so

$$P(\text{spin multiple of 2}) = \frac{3}{6} = \frac{1}{2}$$

3. There is only two outcomes as greater than 4, they are 5 and 6, so

$$P(\text{spin a number greater than 4}) = \frac{2}{6} = \frac{1}{3}$$

4. There is not number that has both values as 2 and 3, so

$$P(\text{spin 2 and 3}) = \frac{0}{6} = 0$$

Note: In Q4, if there is "or" instead of "and", the probability will be different

$$P(\text{spin 2 or 3}) = \frac{2}{6} = \frac{1}{3}$$

#### Example 4.1.9: Find Theoretical Probabilities (Roll of Two Die)

We should think of this experiment as occurring in two stages: (1) one die roll, then (2) another die roll. Even though these two stages will usually occur simultaneously in practice, since they're independent, it's okay to treat them separately. Now, each of the 36 ordered pairs in the table represents an equally likely outcome.

Since we have two independent stages, let's create a table (Figure 4.1.7), which is probably the most efficient method for determining the sample space.

|            |   | First Die |        |        |        |        |        |
|------------|---|-----------|--------|--------|--------|--------|--------|
|            |   | 1         | 2      | 3      | 4      | 5      | 6      |
| Second Die | 1 | (1, 1)    | (2, 1) | (3, 1) | (4, 1) | (5, 1) | (6, 1) |
|            | 2 | (1, 2)    | (2, 2) | (3, 2) | (4, 2) | (5, 2) | (6, 2) |
|            | 3 | (1, 3)    | (2, 3) | (3, 3) | (4, 3) | (5, 3) | (6, 3) |
|            | 4 | (1, 4)    | (2, 4) | (3, 4) | (4, 4) | (5, 4) | (6, 4) |
|            | 5 | (1, 5)    | (2, 5) | (3, 5) | (4, 5) | (5, 5) | (6, 5) |
|            | 6 | (1, 6)    | (2, 6) | (3, 6) | (4, 6) | (5, 6) | (6, 6) |

Figure 4.1.7: Sample Space of Rolling Two Dice

To make our analysis easier, let's replace each ordered pair with the sum, and since the event we're interested in is the one consisting of rolls of 4, 5, or 7. Let's shade those in (Figure 4.1.8.)

|            |   | First Die |   |   |    |    |    |
|------------|---|-----------|---|---|----|----|----|
|            |   | 1         | 2 | 3 | 4  | 5  | 6  |
| Second Die | 1 | 2         | 3 | 4 | 5  | 6  | 7  |
|            | 2 | 3         | 4 | 5 | 6  | 7  | 8  |
|            | 3 | 4         | 5 | 6 | 7  | 8  | 9  |
|            | 4 | 5         | 6 | 7 | 8  | 9  | 10 |
|            | 5 | 6         | 7 | 8 | 9  | 10 | 11 |
|            | 6 | 7         | 8 | 9 | 10 | 11 | 12 |

Figure 4.1.8: Possible Sum When Rolling Two Dice

Our event has 13 outcomes, so the probability that your sister rolls a losing number is  $\frac{13}{36}$ .

**Example 4.1.10: Tree Diagrams to Compute Theoretical Probability (Flip Coins 3 times)**

As we see in figure 4.1.4, the sample space is {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}, which has 8 outcomes.

1. Flipping exactly 2 heads occurs three times (HHT, HTH, THH), so the probability is

$$P(\text{Exactly two heads}) = \frac{3}{8}$$

2. Flipping 2 consecutive heads at some point in the experiment happens 3 times: HHH, HHT, THH. So, the probability is

$$P(\text{Consecutive two heads}) = \frac{3}{8}$$

3. There are 2 outcomes that all show the same result: HHH and TTT. So, the probability is

$$P(\text{All same result}) = \frac{2}{8} = \frac{1}{4}$$

**Example 4.1.11: Finding Empirical Probabilities**

1. Since Jose made 47 out of his last 80 attempts, assign this event an empirical probability of

$$P(\text{Make}) = \frac{47}{80} \approx 59\%$$

2. Amy completed the commute in under 30 minutes in 12 of the last 60 commutes, so we can estimate her probability of making it in under 30 minutes is

$$P(\text{Making under 30 minutes}) = \frac{12}{60} \approx 20\%$$

3. Since Felix has won 14 of the last 20 games, assign a probability for a win this time of

$$P(\text{Win}) = \frac{14}{20} \approx 70\%$$

**Example 4.1.12: Empirical Probability**

1. There are 1,600 M&M. Number of yellow M&M 116. So the probability of drawing yellow is

$$P(\text{Yellow}) = \frac{116}{1,600} = 0.0725 = 7.25\%$$

2. There are  $116 + 419 = 535$  "yellow or green" M&M in a case.

$$P(\text{Yellow or Green}) = \frac{535}{1,600} = 0.3344$$

This page titled 4.1: Fundamental Concepts of Probability is shared under a license and was authored, remixed, and/or curated by David Lippman (The OpenTextBookStore) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.5: Basic Concepts of Probability by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/contemporary-mathematics>.
- 12.2: Basic Concepts by David Lippman is licensed CC BY-SA 3.0. Original source: <http://www.opentextbookstore.com/mathinsociety>.
- 7.4: Tree Diagrams, Tables, and Outcomes by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/contemporary-mathematics>.

## 4.2: The Addition and Complement Rules

### Complementary Events

Now, let us examine the probability that an event does **not** happen. As in the previous section, consider the situation of rolling a 6-sided die and first compute the probability of rolling a 6: the answer is  $P(6) = \frac{1}{6}$ . Now consider the probability that we do *not* roll a 6: there are 5 outcomes that are not a 6, so the answer is  $P(\text{not a 6}) = \frac{5}{6}$ . Notice that

$$P(6) + P(\text{not a 6}) = \frac{1}{6} + \frac{5}{6} = \frac{6}{6} = 1$$

So we have a rule

$$P(\text{not } A) = 1 - P(A)$$

Let's summarize the above discussion.

#### Complement of an Event

The **complement** of event  $A$  is the event " $A$  does not happen (not  $A$ ).\" It is the set of outcomes in the sample space  $S$  that are not in event  $A$ .

- The notation  $A'$  is used for the complement of event  $A$ .
- The probability of the complement of event  $A$  is  $P(A') = 1 - P(A)$ .
- Notice also that  $P(A) = 1 - P(A')$  since event  $A$  and its complement make up the entire sample space,  $P(A) + P(A') = 1$ .
- Note sometime  $A'$  is also written as  $A^c$

Note that the complement of an event in probability is essentially the same as that of a set since an event is a subset of the sample space.

#### Example 4.2.1: Complement of an Event

1. If you pick a random card from a deck of playing cards, what is the probability it is not a heart?
2. ABC Airlines reports that 90.6% of its flights arrive on time. What percent of its flights do not arrive on time?

*Answer available at the end of this section.*

#### Your Turn 4.2.1: Find the Probability of Complement

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Addition Rule for “Or” Probabilities

Many probabilities in real life involve more than one event. If we draw a single card from a deck we might want to know the probability that it is either red or a jack. If we look at a group of students, we might want to know the probability that a single student has brown hair and blue eyes. When we combine two events, we make a single event called a **compound event**. To create a compound event, we can use the word “and” or the word “or” to combine events. It is very important in probability to pay attention to the words “and” and “or” if they appear in a problem. The word “and” restricts the field of possible outcomes to only those outcomes that simultaneously describe all events. The word “or” broadens the field of possible outcomes to those that describe one or more events. This is essentially the same as unions and intersections in sets.

Note that the non-exclusive use of the word “or” will double-count certain events. Say we look at the event of drawing a single card from a deck and the outcome of drawing a “red card or a jack”. If you count the number of options for red cards and the number of options for a jack, you are counting the red jacks twice. This double counting needs to be accounted for.

Suppose a teacher wants to know the probability that a student in her class of 30 is taking either Art or English. She asks the class to raise their hands if they are taking Art and counts 13 hands. Then she asks the class to raise their hands if they are taking English and counts 21 hands. The teacher then calculates (the incorrect answer)

$$P(\text{Art or English}) = \frac{13}{30} + \frac{21}{30} = \frac{13 + 21}{30} = \frac{34}{30} > 1$$

The teacher knows this is wrong because probabilities must be inclusive and between 0 and 1. After thinking about it, she remembered nine students taking both Art and English. These students raised their hands each time she counted, so the teacher counted them twice. When we calculate probabilities, we have to be careful to count each outcome only once.

The correct answer would be:

$$P(\text{Art or English}) = \frac{13 + 21 - 9}{30} = \frac{25}{30} = \frac{5}{6}$$

We will work out how to get the correct answer in general below.

An experiment involves drawing one card from a well-shuffled deck of 52 cards. Consider the events  $E$  = the card is red,  $F$  = the card is a 5, and  $G$  = the card is a spade. A card can be both red and a 5 simultaneously, but it is not possible for a card to be both red and a spade simultaneously. It would be easy to accidentally count a red 5 twice by mistake. It is not possible to count a red spade twice.

If two events have events in common or can happen simultaneously, the overlap is called the **intersection** of the events. The intersection of events is denoted as  $A$  “and”  $B$ , and is the same as the intersection of two sets  $A$  and  $B$ .

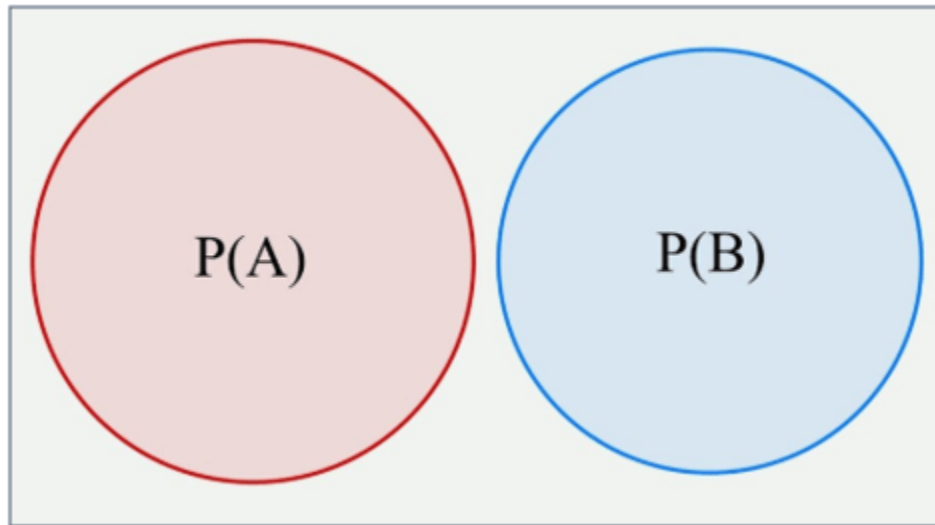


Figure 4.2.1: Mutually exclusive events

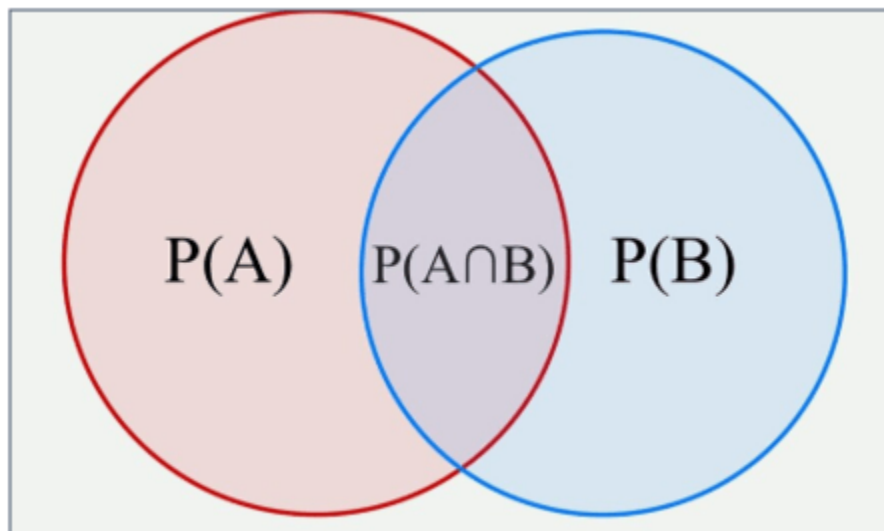


Figure 4.2.2: Not mutually exclusive events

Two events are **mutually exclusive** if they have no outcomes in common. In other words, **mutually exclusive events** are those that **cannot occur simultaneously**.

The probability of two mutually exclusive events  $A$  and  $B$  is  $P(A \text{ and } B) = 0$ .

### Examples of Mutually Exclusive and Not Mutually Exclusive Events

Example of mutually exclusive events:

1. Drawing one card from a deck, A: Drawing a red card, and B: Drawing a black card
2. Rolling a die, A: Rolling a 2 and B: Rolling a 3
3. A student is chosen randomly: A: Student passes the exam, and B: Student fails the exam

Examples of not mutually exclusive events:

1. Rolling a die, A: Rolling an even number, and B: Rolling a number greater than 3
2. Drawing one card from a deck, A: Drawing a red card (hearts or diamonds), and B: Drawing a king
3. A student is chosen randomly. A: Student is in the drama club, and B: Student is in the soccer team

#### Example 4.2.2: Identify if the Events are Mutually Exclusive or not?

Two fair dice are tossed, and different events are recorded. Let the events  $A$ ,  $B$  and  $C$  be as follows:

$A$  = The sum is 5 =  $\{(1, 4), (2, 3), (3, 2), (4, 1)\}$

$B$  = Both numbers are even =  $\{(2, 2), (2, 4), (2, 6), (4, 2), (4, 4), (4, 6), (6, 2), (6, 4), (6, 6)\}$

$C$  = Both numbers are less than 5  
 $= \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4), (4, 1), (4, 2), (4, 3), (4, 4)\}$

- a. Are events  $A$  and  $B$  mutually exclusive?
- b. Are events  $A$  and  $C$  mutually exclusive?
- c. Are events  $B$  and  $C$  mutually exclusive?

*Answer available at the end of this section.*

*Answer available at the end of this section.*

#### Your Turn 4.2.2: Disjoint or Not

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



#### Do Not Count the Same Outcome Twice.

The addition rule for probabilities is used when the word "or" connects the events." Remember, the teacher in her class wanted to know the probability that her students took either Art or English. Her problem was that she counted some students twice. She needed to add the number of students taking Art to the number of students taking English, and then subtract the number of students she counted twice. After dividing the result by the total number of students, she will find the desired probability. The calculation is as follows:

$$\begin{aligned}
 P(\text{Art or English}) &= \frac{\# \text{ taking Art} + \# \text{ taking English} - \# \text{ taking both}}{\text{total number of students}} \\
 &= \frac{13 + 21 - 9}{30} \\
 &= \frac{25}{30} \approx 0.833
 \end{aligned}$$

The probability that a student takes Art or English is 0.833 or 83.3%.

When calculating the probability for compound events connected by the word "or," we must be careful not to count the same thing twice.

If we want the probability of drawing a red card or a 5 we cannot count the red 5s twice. If we want to know the probability of a person being blonde-haired or blue-eyed, we cannot count the blue-eyed blondes twice. The addition rule for probabilities adds the number of blonde-haired people to the number of blue-eyed people and then subtracts the number of people we counted twice.

### Formula: Addition Rule for “Or” Probabilities

If  $A$  and  $B$  are any events, then the probability of either  $A$  or  $B$  occurring (or both) is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If  $A$  and  $B$  are mutually exclusive events then  $P(A \text{ and } B) = 0$ , so then

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

Slightly different way, we can see the above formula

$$P(A \text{ or } B) = \frac{n(A) + n(B) - n(A \text{ and } B)}{\text{total number outcomes}}$$

Where  $n(A)$  = Number of outcomes in  $A$ .

### Example 4.2.3: Deck of Card Mutually Exclusive

A single card is drawn from a well-shuffled deck of 52 cards. Find the probability that the card is a club or a face card.

*Answer available at the end of this section.*

### Your Turn 4.2.3: Find Or Probability

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 4.2.4: Deck of Cards (Mutually Exclusive and Not Mutually Exclusive)

1. Suppose we draw one card from a standard deck. What is the probability that we get a queen or a king?
2. Suppose we draw one card from a standard deck. What is the probability that we will get a red card or a king?

Answer 1 available at the end of this section.

### Your Turn 4.2.4: Find Or Probability (Two Dice Rolled)

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Find the Probability Using a Two-way Table

A **two-way table** (also called a **contingency table**) is a way to organize data that shows the relationship between **two categorical variables**. It helps you see how different categories intersect. It's like a grid with one variable across the **top (columns)** and another variable down the **side (rows)**, and each **cell** shows the **frequency/count** of observations for that pair of categories.

### KEEP IN MIND

When you find a probability using a **two-way table**, you **must always use the total number of observations** (the **grand total**) in the **denominator**, unless the problem specifically tells you to find a *conditional probability*.

### Example 4.2.5: Calculate Probability Using Two Way Table

A survey was conducted among 250 people who had recently purchased a car, and the results are summarized in the following table.

*Satisfaction of Car Buyers*

|          | Satisfied | Not Satisfied | Total |
|----------|-----------|---------------|-------|
| New Car  | 92        | 28            | 120   |
| Used Car | 83        | 47            | 130   |
| Total    | 175       | 75            | 250   |

Find the probability that a person bought a new car or was not satisfied.

Answer available at the end of this section.

### Your Turn 4.2.5: Probability Using Two Way Table

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

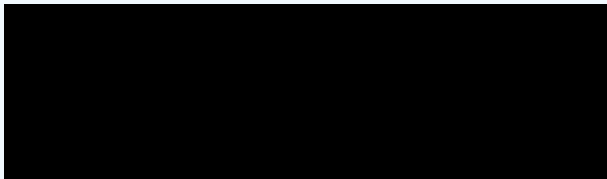
[View this Interactive Content online](#)



### Example 4.2.6: Calculate Probability Using Two Way Table

The table below shows the number of survey subjects who have received and have not received a speeding ticket in the last year, and the color of their car. Find the probability that a randomly chosen person:

1. Has a red car *and* got a speeding ticket
2. Has a red car *or* got a speeding ticket.



Answer available at the end of this section.

### Your Turn 4.2.6: Calculate Probability Using Venn Diagram

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Answers

### Example 4.2.1: Complement of an Event

1. There are 13 hearts in the deck, so

$$P(\text{heart}) = \frac{13}{52} = \frac{1}{4}$$

The probability of *not* drawing a heart is the complement of drawing a heart:

$$P(\text{not heart}) = 1 - P(\text{heart}) = 1 - \frac{1}{4} = \frac{3}{4}$$

2. Let  $A$  = arrive on time. The probability of  $A$  is  $P(A) = 0.906$ . The probability of not arriving on time is

$$P(A') = 1 - P(A) = 1 - 0.906 = 0.094 = 9.4\%$$

### Example 4.2.2: Identify if the Events are Mutually Exclusive or not?

#### Example 4.2.2: Identify if the Events are Mutually Exclusive or not?

- a. Yes.  $A$  and  $B$  are mutually exclusive because they have no outcomes in common. It is not possible to add 2 even numbers to get a sum of 5.
- b. No.  $A$  and  $C$  are not mutually exclusive because they have some outcomes in common. The pairs (1, 4), (2, 3), (3, 2), and (4, 1) all have sums of 5, and both numbers are less than 5.
- c. No.  $B$  and  $C$  are not mutually exclusive because they have some outcomes in common. The pairs (2, 2), (2, 4), (4, 2), and (4, 4) all have two even numbers that are less than 5.

### Example 4.2.3: Deck of Card Mutually Exclusive

There are 3 cards that are clubs, 12 face cards (J, Q, K in each suit), and 3 face cards that are clubs.

$$\begin{aligned} P(\text{club or face card}) &= P(\text{club}) + P(\text{face card}) - P(\text{club and face card}) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= \frac{22}{52} = \frac{11}{26} \approx 0.423 \end{aligned}$$

The probability that the card is a club or a face card is approximately 0.423 or 42.3%.

**OR** you can do the following directly

$$\begin{aligned} P(\text{club card or face card}) &= \frac{\# \text{ of club card} + \# \text{ of face card} - \# \text{ of both club and face card}}{\text{total number cards}} \\ &= \frac{13 + 12 - 3}{52} \\ &= \frac{22}{52} \approx 0.423 \end{aligned}$$

A simple way to check this answer is to take the 52 card deck and count the number of physical cards that are either clubs or face cards. If you were to set aside all of the clubs and face cards in the deck, you would end up with the following:

**OR** you can also do the following

{2 Clubs, 3 Clubs, 4 Clubs, 5 Clubs, 6 Clubs, 7 Clubs, 8 Clubs, 9 Clubs, 10 Clubs, J Clubs, Q Clubs, K Clubs, A Clubs, J Hearts, Q Hearts, K Hearts, J Spades, Q Spades, K Spades, J Diamonds, Q Diamonds, K Diamonds}

That is 22 cards out of the 52 card deck, which gives us a probably of:

$$\frac{22}{52} = \frac{11}{26} \approx 0.423$$

This confirms our earlier answer using the formal Addition Rule.

### Example 4.2.4: Deck of Cards (Mutually Exclusive and Not Mutually Exclusive)

The event of drawing a king card and drawing a queen is mutually exclusive. There are 4 queens and 4 kings in the deck, hence 8 outcomes corresponding to a queen or king out of 52 possible outcomes. Thus, the probability of drawing a queen or a king is:

$$P(\text{King or Queen}) = \frac{8}{52} = \frac{2}{13}$$

Note that in this case, there are no cards that are both a queen and a king, so  $P(\text{King and Queen}) = 0$ . Using the addition rule, we could have said:

$$P(\text{King or Queen}) = P(\text{King}) + P(\text{Queen}) - P(\text{King and Queen}) = \frac{4}{52} + \frac{4}{52} - 0 = \frac{8}{52} = \frac{2}{13}$$

Note that you should not simplify the individual fractions before adding/subtracting them, since they naturally have a common denominator. Simplify after the addition.

**OR** you can do the following directly

$$\begin{aligned} P(\text{King or Queen}) &= \frac{\# \text{ of Kings} + \# \text{ of Queen} - \# \text{ of King and Queen}}{\text{total number of outcomes in sample space}} \\ &= \frac{4 + 4 - 0}{52} \\ &= \frac{8}{52} = \frac{2}{13} \end{aligned}$$

The two events, drawing a red card and drawing a king, are not mutually exclusive. Half the cards are red, so

$$P(\text{red}) = \frac{26}{52} \quad (4.2.1)$$

There are 4 kings, so

$$P(\text{King}) = \frac{4}{52} \quad (4.2.2)$$

There are 2 red kings, so

$$P(\text{red and King}) = \frac{2}{52} \quad (4.2.3)$$

We can then calculate

$$\begin{aligned} P(\text{red or King}) &= P(\text{Red}) + P(\text{King}) - P(\text{Red and King}) \\ &= \frac{26}{52} + \frac{4}{52} - \frac{2}{52} \\ &= \frac{28}{52} = \frac{7}{13} \end{aligned} \quad (4.2.4)$$

Note that you should not simplify the individual fractions before adding/subtracting them, since they naturally have a common denominator. Simplify after the addition.

**OR** you can do the following directly

$$\begin{aligned} P(\text{King or Red}) &= \frac{\# \text{ of King} + \# \text{ of Red} - \# \text{ of King and Red}}{\text{total number of outcomes in sample space}} \\ &= \frac{4 + 26 - 2}{52} \\ &= \frac{28}{52} = \frac{7}{13} \end{aligned}$$

**Example 4.2.5: Calculate Probability Using Two Way Table**

$$\begin{aligned} P(\text{new car or not satisfied}) &= P(\text{new car}) + P(\text{not satisfied}) - P(\text{new car and not satisfied}) \\ &= \frac{120}{250} + \frac{75}{250} - \frac{28}{250} = \frac{167}{250} \approx 0.668 \end{aligned}$$

The probability that a person bought a new car or was not satisfied is approximately 0.668 or 66.8%.

**OR** you can find the probability directly as shown below

$$\begin{aligned} P(\text{new car or not satisfied}) &= \frac{\# \text{ of new car} + \# \text{ of not satisfied} - \# \text{ of new are and not satisfied}}{\text{total number people}} \\ &= \frac{120 + 75 - 28}{250} \\ &= \frac{167}{250} \approx 0.668 \end{aligned}$$

**Example 4.2.6: Calculate Probability Using Two Way Table**

1. We can see that 15 people of the 665 surveyed had both a red car and got a speeding ticket, so the probability is

$$P(\text{red car and ticket}) = \frac{15}{665} \approx 0.0226 \quad (4.2.5)$$

Notice that having a red car and getting a speeding ticket are not independent events, so the probability of both of them occurring is not simply the product of the probabilities of each one occurring.

2. Of the 665 surveyed, we can see that 150 people had red cars, 60 people had speeding tickets, and 15 people had red cars and speeding tickets. So

$$\begin{aligned} P(\text{red car or ticket}) &= \frac{\# \text{ of red car} + \# \text{ of ticket} - \# \text{ of red car and got ticket}}{\text{total number people surveyed}} \\ &= \frac{150 + 60 - 15}{665} \\ &= \frac{195}{665} = \frac{39}{133} \approx 0.2932 \end{aligned}$$

---

This page titled [4.2: The Addition and Complement Rules](#) is shared under a license and was authored, remixed, and/or curated by [David Lippman](#) ([The OpenTextBookStore](#)) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.3: Working with Events](#) by [David Lippman](#) is licensed [CC BY-SA 3.0](#). Original source: <http://www.opentextbookstore.com/mathinsociety>.
- [3.2: Combining Probabilities with “And” and “Or”](#) by [Maxie Inigo](#), [Jennifer Jameson](#), [Kathryn Kozak](#), [Maya Lanzetta](#), and [Kim Sonier](#) is licensed [CC BY-SA 4.0](#). Original source: <https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier>.
- [3.3: Conditional Probabilities](#) by [Maxie Inigo](#), [Jennifer Jameson](#), [Kathryn Kozak](#), [Maya Lanzetta](#), and [Kim Sonier](#) is licensed [CC BY-SA 4.0](#). Original source: <https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier>.

## 4.3: Independence and Multiplication Rules

In probability, **independent events** are events where the occurrence of one does not affect the occurrence of the other. For example, flipping a coin and rolling a die are independent events because the outcome of the coin flip (heads or tails) does not influence the number that appears on the die. Moreover, when selecting a **small sample from a large population**, individual selections can often be treated as **independent events** because the removal of one item does not significantly alter the probabilities of selecting subsequent items. This is particularly true when the sample size is **much smaller** than the total population.

### Dependent Vs Independent Events

|            | Independent Events                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Dependent Events                                                                                                                                                                                                                                                                                                                                                                                         |
|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Definition | If the occurrence of one event has <b>no effect</b> on the probability of the other. The "pool" of possibilities remains the same across trials.                                                                                                                                                                                                                                                                                                                       | If the outcome of the first event <b>changes</b> the probability of the next. This usually happens when the sample space is altered.                                                                                                                                                                                                                                                                     |
| Examples:  | <ol style="list-style-type: none"> <li>1. Flipping a coin and then flipping the coin again.</li> <li>2. Lacy draws a card from a standard deck of 52 cards. She returns to the deck, then draws a second card.</li> <li>3. A bag contains 6 yellow blocks and 4 purple blocks. Susan picks a block from the bag and then another block without replacing the first one.</li> <li>4. Even A: It rains tomorrow, and Event B: You get an A on your math test.</li> </ol> | <ol style="list-style-type: none"> <li>1. A bag contains 6 yellow blocks and 4 purple blocks. Susan picks a block from the bag and then another block without replacing the first one.</li> <li>2. Lacy draws a card from a standard deck of 52 cards. She returns to the deck, then draws a second card.</li> <li>3. Event A: snows tomorrow, and Event B: School will be canceled tomorrow.</li> </ol> |

### Your Turn 4.3.1: Identify Independent Event

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Definition: Multiplication Rule

When two events A and B are independent, then

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

This formula can be extended to any number of independent events, then

$$P(A \text{ and } B \text{ and } C) = P(A) \cdot P(B) \cdot P(C)$$

### Example 4.3.2: Using Multiplication Rule

Assume that two dice are tossed twice. Compute the following probabilities. Write the answers as a fraction or a decimal rounded to three places.

1. Get a sum of 7 on both tosses.
2. Get a sum of 5 on the first toss and a sum of 8 on the second toss.

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

#### Your Turn 4.3.2: Using Multiplication Rule

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



#### Example 4.3.3: Using Multiplication Rule

What is the probability of obtaining five heads in a row when flipping a fair coin? Interpret this probability.

Answer available at the end of this section.

#### Your Turn 4.3.3: Using Multiplication Rule

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



#### Example 4.3.4: Using Multiplication Rule

Three scholarships of \$500 are to be awarded to three STEM majors. The STEM majors who qualify for the scholarship consist of 7 physics majors, 4 math majors, 3 computer science majors, and 6 biology majors. Compute the chance that all three scholarships will be awarded to the math majors. Write the answer as a decimal rounded to three non-zero decimal places.

Answer available at the end of this section.

#### Your Turn 4.3.4: Identify Independent Event

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## At Least One Rule for Probabilities

### How to Use at least one Rule

The probability of **"at least one"** refers to the likelihood that an event occurs at least once in multiple trials. It means that the event happens one or more times, as opposed to not happening at all.

$$P(\text{at least one event A in } n \text{ trials}) = 1 - P(\text{no A in } n \text{ trials})$$

or

$$P(\text{at least one event A in } n \text{ trials}) = 1 - (\text{not A})^n$$

To find **"at least one"** probability, look for the following:

1. Find how many trials for that event (Or find the value of  $n$ )
2. Find the probability of not an event.
3. Use the given formula.

### Example 4.3.5: Using at Least One Rule

The chance that a person is a "night owl" is around 12%. Night owls are those people who believe they are more productive in the late hours of the night. If a sample of 3 people is selected from the population, compute the following probabilities.

1. All three are night owls.
2. None of the three is a night owl.
3. At least one of the three is a night owl.

*Answer available at the end of this section.*

### Your Turn 4.3.5: Multiplication Rule

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 4.3.6: Using Multiplication Rule

A student takes 3 multiple-choice quiz questions. Each question has 4 answer choices, and the student randomly guesses on each question. Each question is independent. What is the probability that the student gets at least one question correct?

*Answer available at the end of this section.*

## Your Turn 4.3.6: Multiplication Rule

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Author/Attribute

"4.4: Multiplication Rules and Conditional Probability"(opens in new window) by Toros Berberyan(opens in new window), Tracy Nguyen(opens in new window), and Alfie Swan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

"4.3: Conditional Probability"(opens in new window) by Kathryn Kozak(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

## Answers

### Example 4.3.2: Using Multiplication Rule

Since tossing two dice is an independent event. Use the multiplication rule to find the probability.

$$P(\text{First is a 7 and Second is a 7}) = \frac{6}{36} \cdot \frac{6}{36} = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 0.028$$

### Example 4.3.2: Using Multiplication Rule

$$P(\text{First is a 5 and Second is an 8}) = \frac{4}{36} \cdot \frac{5}{36} = \frac{1}{9} \cdot \frac{5}{36} = \frac{5}{324} = 0.015$$

### Example 4.3.3: Using Multiplication Rule

Note:  $P(\text{Head}) = \frac{1}{2}$

$P(\text{head in first time and head in second time and head in third time and head in fourth time and head in fifth time})$ .

$$\begin{aligned} P(5 \text{ heads in a row}) &= P(H) \times P(H) \times P(H) \times P(H) \times P(H) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \\ &= \left(\frac{1}{2}\right)^5 = 0.03125 \end{aligned}$$

**Interpretation:** If we flipped a coin 100 times, we would expect to observe 5 heads in a row about  $3 = 0.03125 \times 100$  times.

### Example 4.3.4: Using Multiplication Rule

Since a person can only win the award once, the events are dependent, and Multiplication is applied to solve the answer.

$$P(\text{All 3 are Math Majors}) = \frac{4}{20} \times \frac{3}{19} \times \frac{2}{18} = \frac{1}{285} = 0.00351$$

### Example 4.3.5: Using at Least One Rule

1. These events are independent. We can use the multiplication rule to compute the answer.

$$P(\text{all three are night owls}) = 0.12 \times 0.12 \times 0.12 = (0.12)^3 = 0.0017$$

2. The complement must be computed first. It is  $P(\text{not a night owl}) = 1 - 0.12 = 0.88$ .

$$P(\text{all three are not night owls}) = 0.88 \times 0.88 \times 0.88 = (0.88)^3 = 0.681 \text{ \textcolor{red}{nonnumber}} \quad (4.3.1)$$

3. We use at least one rule and the result of part 2 to get the answer.  $P(\text{At least one is a night owl}) = 1 - 0.681 = 0.319$ .

$$\begin{aligned} P(\text{at least one night owl in 3}) &= 1 - (\text{not night owl})^3 \\ &= 1 - 0.88^3 \\ &= 0.03185 \end{aligned}$$

### Example 4.3.6: Using Multiplication Rule

First, calculate the probability of guessing correctly. It is

$$P(\text{correct}) = \frac{1}{4}$$

Second, calculate the probability of guessing incorrectly. It is

$$P(\text{incorrect}) = 1 - \frac{1}{4} = \frac{3}{4}$$

Third, calculate the probability that all three are guessed incorrectly. It is

$$P(\text{all three are incorrect}) = \left(\frac{3}{4}\right)^3 = \frac{27}{64}$$

Finally, calculate the probability that at least one problem is guessed correctly.

$$\begin{aligned}P(\text{at least one guessed correctly out of 3}) &= 1 - [P(\text{none correctly})]^3 \\&= 1 - \left(\frac{3}{4}\right)^3 \\&= 1 - \frac{27}{64} \\&= \frac{37}{64}\end{aligned}$$

---

This page titled [4.3: Independence and Multiplication Rules](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.1: Prelude to Exponential and Logarithmic Functions](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.

## 5: DISCRETE PROBABILITY DISTRIBUTIONS

5.1: Discrete Probability Distributions

5.2: Binomial Probability Distributions

---

This page titled [5: Discrete Probability Distributions](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Kathryn Kozak](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [11.5: Hyperbolas](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.

## 5.1: Discrete Probability Distributions

### Random Variable

As a reminder, a variable, or what will be called the random variable from now on, is represented by the letter  $X$ , and it represents a quantitative (numerical) variable that is measured or observed in an experiment. In other words, a **random variable** is a rule that assigns a single numerical value to each outcome of a probability experiment. It converts outcomes—like words, symbols, or measurements—into numbers so you can analyze them mathematically.

Consider an experiment: tossing a coin twice.

- Outcomes: Heads-Heads, Heads-Tails, Tails-Heads, Tails-Tails
- Random Variable  $X$ : Count the total number of Heads.

#### Definition: Random Variable

A random variable, usually written  $X$ , is a function that assigns a real number to each outcome in the sample space of a random experiment. The lower case  $x$  represents the possible values of the random variables.

Also, remember that quantitative variables are classified as discrete or continuous. What is the difference between discrete and continuous data? **Discrete** data can only take on particular values in a range. **Continuous** data can take on any value in a range. Discrete data usually arises from counting, while continuous data usually arises from measuring. If you have a variable that can be assigned a probability, it is called a **random variable**. In many cases, the random variable is what you are measuring, but when it comes to discrete random variables, it is usually what you are counting

#### Example of Random Variables

| Experiment                                              | Random Variable and its Type              | The values of Random Variables |
|---------------------------------------------------------|-------------------------------------------|--------------------------------|
| Ask the basketball player to do three free throws       | $X$ : number of shots he makes (Discrete) | $x = 0, 1, 2, 3$               |
| Count absentees in a class with 30 students             | $X$ : number of absentees (Discrete)      | $x = 0, 1, 2, \dots, 30$       |
| Measure how much rain in the month of July              | $X$ : amount of rain value (Continuous)   | $x > 0$                        |
| Measure the height of a plant with the given fertilizer | $X$ : height of a plant (Continuous)      | $x > 0$                        |

How tall is a plant given a new fertilizer? Continuous. This is something you measure. How many fleas are on prairie dogs in a colony? Discrete. This is something you count.

#### Your Turn 5.1.1: Identify the Type of a Variables

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



# Probability Distribution

Now, suppose you put all the values of the random variable together with the probability that it would occur. You could then have a distribution like before, but now it is called a probability distribution since it involves probabilities. A **probability distribution** is an assignment of probabilities to the values of the random variable. A **probability distribution** is a rule, table, or mathematical formula that lists all possible values of a random variable along with how likely each value is to occur.

## Rule for Probability Distribution

A probability distribution is a list or table that shows each possible value of the variable and the probability of that value occurring. The following conditions must be satisfied for a probability distribution.

1.  $0 \leq P(x) \leq 1$
2.  $\sum P(x) = 1$

## Example of Probability Distribution

Let  $X$  be the number of heads obtained when two coins are tossed.

$$S = \{HH, HT, TH, TT\}$$

The possible value of  $X$  are,  $x = 0, 1, 2$

$x = 0$  (no heads  $\rightarrow TT$ )

$x = 1$  (one head  $\rightarrow HT, TH$ )

$x = 2$  (two heads  $\rightarrow HH$ )

Table 1: Probability Distribution Table

| $X$    | 0             | 1             | 2             |
|--------|---------------|---------------|---------------|
| $P(x)$ | $\frac{1}{4}$ | $\frac{2}{4}$ | $\frac{1}{4}$ |

This is probability distribution because,  $0 \leq P(x) \leq 1$  and  $P(0) + P(1) + P(2) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1$

## Example 5.1.2: Probability Distribution

The 2010 U.S. Census found the chance of a household being a certain size. The data is in table 5.1.1 ("Households by age," 2013).

1. Verify if it is a probability distribution.
2. Graph the given table.
3. Find  $P(x = 5)$ .
4. Find  $P(x \leq 3)$ .

Table 5.1.2: Household Size from US Census of 2010

| Size of household | 1     | 2     | 3     | 4     | 5    | 6    | 7 or more |
|-------------------|-------|-------|-------|-------|------|------|-----------|
| Probability       | 26.7% | 33.6% | 15.8% | 13.7% | 6.3% | 2.4% | 1.5%      |

Answer available at the end of this section.

### Your Turn 5.1.2: Probability Distribution

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Mean and Standard Deviation of a Discrete Probability Distribution

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 2, 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. In other words, after conducting many trials of an experiment, you would expect this average value.

### Mean and Standard Deviation of Probability Distribution

For a discrete probability distribution function,

The mean or expected value

$$\mu_{\bar{x}} = \sum xP(x)$$

The standard deviation

$$\sigma_{\bar{x}} = \sqrt{\sum (x - \mu)^2 P(x)}$$

The variance is

$$\sigma_{\bar{x}}^2 = \sum (x - \mu)^2 P(x)$$

where  $x$  = the value of the random variable and  $P(x)$  = the probability corresponding to a particular  $x$  value.

### Example 5.1.3: Calculating mean, Variance, and Standard deviation for a Discrete Probability Distribution

The 2010 U.S. Census found the chance of a household being a certain size. The data is in the table ("Households by age," 2013).

Table 5.1.3: Household Size from US Census of 2010

| Size of household | 1     | 2     | 3     | 4     | 5    | 6    | 7 or more |
|-------------------|-------|-------|-------|-------|------|------|-----------|
| Probability       | 26.7% | 33.6% | 15.8% | 13.7% | 6.3% | 2.4% | 1.5%      |

- Find the mean, variance, and standard deviation.
- Find mean, variance, and standard deviation using technology

Answer available at the end of this section.

### Your Turn 5.1.3: Mean, Variance, and Standard Deviation of Probability Distribution

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Expected Value

The casino game roulette has dozens of different bets that can be made. These bets have different probabilities of winning but also have different payouts. In general, the lower the probability of winning a bet, the more money a player wins for that bet. With so many options, is there one bet that's "smarter" than the rest? What's the best play to make at a roulette table? This section will develop the tools we need to answer these questions.

Many experiments have numbers associated with their outcomes. Some are easy to define; if you roll 2 dice, the sum of the numbers showing is a good example. In some card games, cards have different point values associated with them; for example, in some forms of the game rummy, aces are worth 15 points; 10 s, jacks, queens, and kings are worth 10; and all other cards are worth 6. The outcomes of casino and lottery games are all associated with an amount of money won or lost. These outcome values are used to find the expected value of an experiment: the mean of the values associated with the outcomes that we would observe over a large number of repetitions of the experiment.

That definition is a little vague; How many is "a large number?" In practice, it depends on the experiment; the number must be large enough that every outcome is expected to appear at least a few times. For example, if we're rolling a standard 6-sided die and noting the number that shows, a few dozen replications should be sufficient to ensure the mean is representative. Since the probability of each outcome is  $\frac{1}{6}$ , we would expect to see each outcome about 8 times over the course of 48 replications. However, if we're talking about the Powerball lottery, where the probability of winning the jackpot is about  $\frac{1}{292,000,000}$ , we would need several *billion* replications to ensure that every outcome appears a few times. Luckily, we can find the theoretical expected value before we even run the experiment for the first time.

### FORMULA: Expected Value

The mean can be thought of as the **expected value**. It is the value you expect to get if the trials were repeated an infinite number of times. The mean or expected value does not need to be a whole number, even if the possible values of  $x$  are whole numbers. Expected value is denoted by  $E(X)$ .

$$\mu_{\bar{x}} = E(X)$$

If  $x$  represents an outcome of an experiment and  $P(x)$  represents the corresponding probability of the outcome. The expected value of the experiment is:

$$\Sigma(x) \times P(x)$$

where  $\Sigma$  is the “sum,” meaning we add up the results of the formula that follows over all possible outcomes.

### Key Points to Remember for Finding Expected Value

1. Identify all possible outcomes and their values. These values can be positive or negative depending on the situation.
2. Determine the probability of each outcome.
3. Set up a probability distribution table:
  - First column (or row) → outcome probabilities
  - Second column (or row) → corresponding outcome numerical value
4. **Multiply each outcome value by its probability, then add all the products together to find the expected value.**

Here are the most common areas where people use expected value:

**Gambling and Games:** Used to determine the average gain or loss in activities like lottery, cards, dice, and slot machines, and to decide if a game is fair or profitable.

**Insurance:** Insurance companies use expected value to set premiums by estimating the expected cost of accidents, health claims, or property damage. Finance and Investments: Used to evaluate risky investments and predict the average profit or loss over time.

**Sports:** Helps predict average points, wins, or scoring outcomes in a game.

Weather and Natural Events: Used to estimate expected rainfall, average temperature, or the likelihood of natural disasters.

### Example 5.1.4: Finding Expected Values

Find the expected values of the following experiments.

An **unfair** six-sided die has sides numbered 1, 2, 3, 4, 5, 6 with probabilities given in the table below. Find the expected value of one roll of the die.

Table 5.1.7: Outcomes for rolling a six-sided die and their probabilities

| Value of Outcome | 1             | 2             | 3             | 4             | 5             | 6             |
|------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Probability      | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{8}$ |

Answer available at the end of this section.

### Your Turn 5.1.4: Finding Expected Value

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example 5.1.5 Calculating the expected value

In the Florida lottery called Pick 3, a player pays \$1 and then picks a three-digit number. If those three numbers are picked in that specific order the person wins \$500. What is the expected value in this game?

Answer available at the end of this section.

### Your Turn 5.1.5: Find the Expected Value

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Determining if an event is unusual

If you are looking at a value of  $x$  for a discrete variable, and the  $P$  (the variable has a value of  $x$  or more)  $< 0.05$ , then you can consider  $x$  an unusually high value. Another way to think of this is if the probability of getting such a high value is less than 0.05, then the event of getting the value  $x$  is unusual.

### Unusual Outcomes

Similarly, if the  $P$  (the variable has a value of  $x$  or less)  $< 0.05$ , then you can consider this an unusually low value. Another way to think of this is if the probability of getting a value as small as  $x$  is less than 0.05, then the event  $x$  is considered unusual.

Why is it " $x$  or more" or " $x$  or less" instead of just " $x$ " when you are determining if an event is unusual?

When deciding whether an event is **unusual**, we are interested in how likely it is to get a result **at least as extreme as the observed value**, not just the probability of getting exactly that value.

For example, if a friend wins 6 **out of** 10 games, the probability of exactly 6 wins is about 20.5%, while the probability of 6 **or more wins** is about 37.7%. The second probability is more appropriate because it includes all outcomes that are at least as unusual in the direction we are investigating.

With a larger sample, this becomes even more important. If a friend wins 510 **out of** 1000 games, the probability of exactly 501 wins may be only about 2.5% but the probability of 501 **or more wins** is about 47.8%. Clearly, getting 501 wins is not unusual.

**Key idea:** When determining whether an observation is unusual, consider the probability of getting the observed value **or something more extreme**, rather than the probability of getting exactly that value.

### Example 5.1.6: Is the Outcome Unusual?

The 2010 U.S. Census found the chance of a household being a certain size. The data is in the table ("Households by age," 2013).

Table 5.1.9: Household Size from US Census of 2010

| Size of household | 1    | 2    | 3    | 4    | 5   | 6   | 7 or more |
|-------------------|------|------|------|------|-----|-----|-----------|
| Probability       | 26.7 | 33.6 | 15.8 | 13.7 | 6.3 | 2.4 | 1.5       |

1. Is it unusual for a household to have six people in the family?
2. If you came upon many families with six people, what would you think?
3. Is it unusual for a household to have four people in the family?
4. If you came upon a family of four, what would you think?

## Answers

### Example 5.1.2: Probability Distribution

In this case, the random variable is  $X$  = number of people in a household. This is a discrete random variable, since you are counting the number of people in a household.

This is a probability distribution since you have the  $x$  values and the corresponding probabilities; all probabilities are between zero and one, and the sum of all probabilities is one.

You draw a histogram with the  $x$  values on the horizontal axis corresponding to the size of the household, and the probabilities are on the vertical axis.

You can also give a probability distribution in table form (as shown in table 5.1.2) or as a graph (Figure 5.1.1). The graph looks like a histogram.

A probability distribution is basically a relative frequency distribution based on a very large sample.

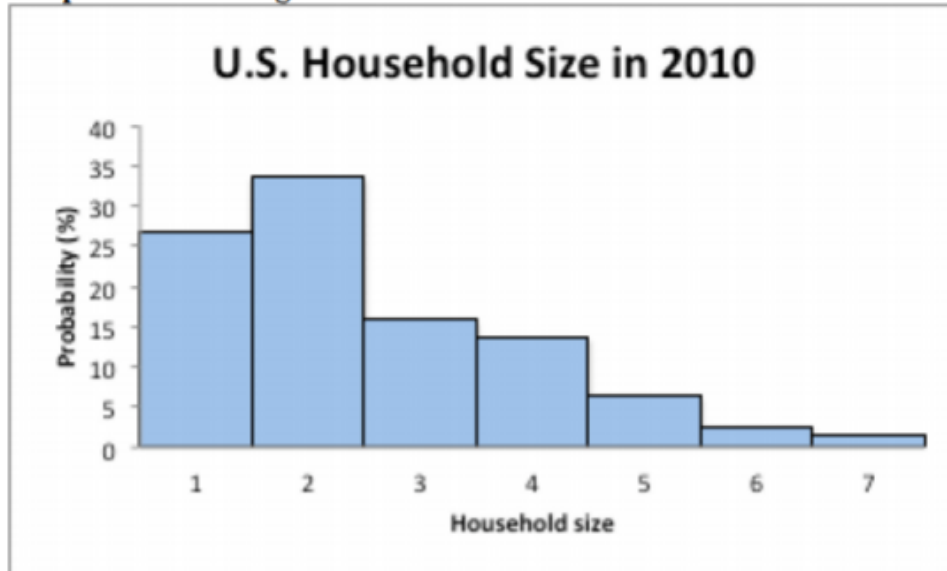


Figure 5.1.1: Histogram of Household Size from US Census of 2010

$P(x = 5) = 0.063$  and  $P(x \leq 3) = 0.267 + 0.336 + 0.158 = 0.761$ .

### Example 5.1.3: Calculating mean, Variance, and Standard deviation for a Discrete Probability Distribution

State random variable:

$X$  = number of people in a household

a. To find the mean it is easier to just use a table as shown below. Consider the category 7 or more to just be 7. The formula for the mean says to multiply the  $x$  value by the  $P(x)$  value, so add a row to the table for this calculation. Also convert all  $P(x)$  to decimal form.

Table 5.1.4: Calculating the Mean for a Discrete PDF

| $x$     | 1     | 2     | 3     | 4     | 5     | 6     | 7     |
|---------|-------|-------|-------|-------|-------|-------|-------|
| $P(x)$  | 0.267 | 0.336 | 0.158 | 0.137 | 0.063 | 0.024 | 0.015 |
| $xP(x)$ | 0.267 | 0.672 | 0.474 | 0.548 | 0.315 | 0.144 | 0.098 |

Now add up the last row, and you get the answer 2.525. This is the mean of a discrete random variable, or the expected value,

$$\mu_x = 2.525$$

This means that you expect a household in the U.S. to have  $2.525 \approx 3$  people in it.

Now, of course, you can't have half a person, but what this tells you is that you expect a household to have either 2 or 3 people, with a little more 3-person households than 2-person households.

To find the variance, again, it is easier to use a table version than try to use the formula in a line. Looking at the formula, you will notice that the first operation that you should do is to subtract the mean from each  $x$  value. Then you square each of these values. Then you multiply each of these answers by the probability of each  $x$  value. Finally, you add up all of these values.

Table 5.1.5: Calculating the Variance for a Discrete PDF

| $x$     | 1     | 2     | 3     | 4     | 5     | 6     | 7     |
|---------|-------|-------|-------|-------|-------|-------|-------|
| $xP(x)$ | 0.267 | 0.336 | 0.158 | 0.137 | 0.063 | 0.024 | 0.015 |

|                              |        |        |        |        |        |         |         |
|------------------------------|--------|--------|--------|--------|--------|---------|---------|
| $x - \mu_{\bar{x}}$          | -1.525 | -0.525 | 0.475  | 1.475  | 2.475  | 3.475   | 4.475   |
| $(x - \mu_{\bar{x}})^2$      | 2.3256 | 0.2756 | 0.2256 | 2.1756 | 6.1256 | 12.0756 | 20.0256 |
| $(x - \mu_{\bar{x}})^2 P(x)$ | 0.6209 | 0.0926 | 0.0356 | 0.2981 | 0.3859 | 0.2898  | 0.3004  |

Now add up the last row to find the variance,

$$\sigma_{\bar{x}}^2 = 2.02375 \text{ people}^2$$

(Note: try not to round your numbers too much so you aren't creating rounding error in your answer. The numbers in the table above were rounded off because of space limitations, but the answer was calculated using many decimal places.)

To find the standard deviation, just take the square root of the variance,

$$\sigma_{\bar{x}} = \sqrt{2.02375} \approx 1.422454$$

This means that you can expect a U.S. household to have 2.525 people in it, with a standard deviation of 1.42 people.

In the online Statistics Calculator, do the following steps

1. Click List Editor
2. Put  $x$  data in L1 and  $P(x)$  data in L2 in order
3. Click Clac and click Single Var Stat
4. Type 1-Var\_stats(L1,L2) and click calculate.

$\bar{x}$  is mean and  $\sigma$  is standard deviation. You square the standard deviation to find the variance.

#### Example 5.1.4: Finding Expected Values

We multiply each value by its probability and add them.

$$\begin{aligned} \text{Expected value} &= 1 \times \frac{1}{8} + 2 \times \frac{1}{8} + 3 \times \frac{1}{8} + 4 \times \frac{1}{4} + 5 \times \frac{1}{4} + 6 \times \frac{1}{8} \\ &= \frac{1}{8} + \frac{2}{8} + \frac{3}{8} + \frac{4}{4} + \frac{5}{4} + \frac{6}{8} \\ &= \frac{30}{8} = 3.75 \end{aligned}$$

So, the expected value of one roll of the dice is 3.750.

#### Example 5.1.5 Calculating the expected value

To find the expected value, you need to first create the probability distribution. In this case, the random variable  $x$  = winnings. If you pick the right numbers in the right order, then you win \$500, but you paid \$1 to play, so you actually win \$499. If you didn't pick the right numbers, you lose the \$1, the  $x$  value is  $-\$1$ . You also need the probability of winning and losing. Since you are picking a three-digit number, and for each digit there are 10 numbers you can pick, each independent of the others, you can use the multiplication rule. To win, you have to pick the right numbers in the right order. The first digit, you pick 1 number out of 10, the second digit you pick 1 number out of 10, and the third digit you pick 1 number out of 10.

The probability of picking the right number in the right order is

$$\frac{1}{10} * \frac{1}{10} * \frac{1}{10} = \frac{1}{1000} = 0.001$$

The probability of losing (not winning) would be

$$1 - \frac{1}{1000} = \frac{999}{1000} = 0.999$$

Putting this information into a table will help to calculate the expected value.

Table 5.1.8: Finding Expected Value

| Win or lose | $x$    | $P(x)$ | $xP(x)$    |
|-------------|--------|--------|------------|
| Win         | \$499  | 0.001  | \$0.499    |
| Lose        | $-\$1$ | 0.999  | $-\$0.999$ |

Now add the two values together to get the expected value. It is

$$\$0.499 + (-\$0.999) = -\$0.50$$

In the long run, you will expect to lose \$0.50.

Since the expected value is not 0, then this game is not fair. Since you lose money, Florida makes money, which is why they have the lottery.

#### Example 5.1.6: Is the Outcome Unusual?

$X$  = number of people in a household

To determine this, you need to look at probabilities. However, you cannot just look at the probability of six people. You need to look at the probability of  $x$  being six or more people or the probability of  $x$  being six or fewer people. The

$$\begin{aligned} P(x \leq 6) &= P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) + P(x = 5) + P(x = 6) \\ &= 26.7\% + 33.6\% + 15.8\% + 13.7\% + 6.3\% + 2.4\% \\ &= 98.5\% \end{aligned}$$

Since this probability is more than 5%, then six is not an unusually low value.

$$\begin{aligned} P(x \geq 6) &= P(x = 6) + P(x \geq 7) \\ &= 2.4\% + 1.5\% \\ &= 3.9\% \end{aligned}$$

Since this probability is less than 5%, then six is an unusually high value. It is unusual for a household to have six people.

Since it is unusual for a family to have six people in it, you may think that either the size of families is increasing from what it was or that you are in a location where families are larger than in other locations.

To determine this, you need to look at probabilities. Again, look at the probability of  $x$  being four or more, or the probability of  $x$  being four or less. The

$$\begin{aligned}P(x \geq 4) &= P(x = 4) + P(x = 5) + P(x = 6) + P(x = 7) \\&= 13.7\% + 6.3\% + 2.4\% + 1.5\% \\&= 23.9\%\end{aligned}$$

Since this probability is more than 5%, four is not an unusually high value. The

$$\begin{aligned}P(x \leq 4) &= P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) \\&= 26.7\% + 33.6\% + 15.8\% + 13.7\% \\&= 89.8\%\end{aligned}$$

Since this probability is more than 5%, four is not an unusually low value. Thus, four is not an unusual size for a family.

Since it is not unusual for a family to have four members.

---

This page titled [5.1: Discrete Probability Distributions](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Kathryn Kozak](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.6: Solve Exponential and Logarithmic Equations](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.
- [7.11: Expected Value](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/contemporary-mathematics>.

## 5.2: Binomial Probability Distributions

### Binomial Experiment

Section 5.1 introduced the concept of a probability distribution. The focus of the section was on discrete **probability distributions** (PDF). To find the pdf for a situation, you usually need to actually conduct the experiment and collect data. Then you can calculate the experimental probabilities. Normally, you cannot calculate the theoretical probabilities instead. However, there are certain types of experiments that allow you to calculate the theoretical probability. One of those types is called a **Binomial Experiment**.

If you know you have a binomial experiment, then you can calculate binomial probabilities. This is important because binomial probabilities come up often in real life. Examples of binomial experiments are:

- Toss a fair coin ten times, and find the probability of getting two heads.
- Twenty people in class, and look for the probability of more than half being women.
- Shoot five arrows at a target, and find the probability of hitting it five times.

#### Criteria for Experiment to Become Binomial

To determine whether a scenario qualifies as a **binomial experiment**, the situation must meet four strict criteria. You can remember these using the acronym **BINS**.

1. There are only two outcomes: success and failure. (**B**inary Outcomes)
2. The  $n$  trials are independent, which means that what happens on one trial does not influence the outcomes of other trials. (**I**ndependent Trials)
3. Fixed number of trials,  $n$ , which means that the experiment is repeated a specific number of times. (**N**umber of Trials)
4. The probability of a success doesn't change from trial to trial, where  $p$  = probability of success and  $1 - p$  = probability of failure. (**S**ame probability of success)

If the experiment is binomial, the random variable  $X$  associated with the experiment is called a binomial random variable.

#### Example 5.2.1: Check if the Experiment is Binomial or Not?

What is the random variable in the following experiment, and also check if the experiment is binomial or not? Why or why not?

1. Flipping a coin 10 times and counting the number of heads.
2. Given a 15-question multiple-choice test. Each question has 4 responses and only one is correct. And the number of correct answers is recorded.
3. Drawing 5 cards from a deck **without replacement** and counting the number of hearts.
4. Given a 10-question test where 5 questions are multiple-choice, and 5 are True/False questions. Each multiple-choice question has 5 responses, and only one is correct. And the number of correct answers is recorded.

*Answer 1 available at the end of this section.*

## Your Turn 5.2.1: Binomial Experiment or Not?

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Probability using the Binomial Formula (or Calculator)

When working with a binomial experiment, questions often ask for the probability of a range of outcomes rather than a single, exact number. This is where **inequalities** come into play.

### Notation: Working When You Find the Probability Using the Binomial

$n$ : Total number of trials

$x$ : number of successes you want

$p$ : probability of success on a single trial

Be careful, success is not always a good thing. Sometimes a success is something that is bad, like finding a defect. A success just means you observed the outcome you wanted.

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  $X$  = is the number of students who withdraw from the randomly selected elementary physics class.

### Definition: Inequality that involves the Binomial Probability Distribution

In binomial probability problems, inequalities are often used to describe probabilities involving the random variable  $X$  where  $X$  represents the number of successes.

*Inequality that involves the Binomial Probability Distribution*

| Phrase                                                   | Math symbol | Phrase                                                  | Math symbol       |
|----------------------------------------------------------|-------------|---------------------------------------------------------|-------------------|
| "at least" or "no less than", "greater than or equal to" | $x \geq a$  | "fewer than" or "less than"                             | $x < a$           |
| "more than" or "greater than"                            | $x > a$     | "no more than" or "at most", or "less than or equal to" | $x \leq a$        |
| "Exactly" or "equal to"                                  | $x = a$     | "Between"                                               | $a \leq x \leq b$ |

### Example 5.2.2: Binomial Experiment Terminology

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. Note that we are interested in the number of students who do the homework.

1. What is a random variable here, and check whether this experiment is binomial or not.

2. What values does  $x$  take on?
3. What is a "failure," in words?
4. The words "at least" translate as what kind of inequality for the probability question?

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

Answer 3 available at the end of this section.

Answer 4 available at the end of this section.

### Your Turn 5.2.2: Terminology of Binomial Experiment

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Formula to Find the Probability of Binomial Distribution

The binomial formula for the probability of  $r$  successes in  $n$  trials is

$$P(x = r) = {}_n C_r p^r q^{n-r} \text{ where } {}_n C_r = \frac{n!}{r!(n-r)!}$$

The  ${}_n C_r$  is the number of combinations of  $n$  things taking  $r$  at a time. It is read " $n$  choose  $r$ ". Some other common notations for  $n$  choose  $r$  are  $C_{n,r}$ , and  $\binom{n}{r}$ .

$n!$  means you are multiplying

$$n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$$

As an example,

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

**NOTE:** We do use an online calculator to compute the probability under a binomial distribution.

### Caution!

When solving problems involving binomial, make sure you define your random variable and state what the values for  $n$ ,  $p$ , and  $x$  are. Also, you need to know the inequality (one of those mentioned above) that is used to describe probabilities.

Without doing this, the problems are much harder.

Watch the attached video to learn how to find probability using an online statistics calculator for binomial random variables.

### Example 5.2.3: Calculating Binomial Probabilities (Using Formula)

When looking at a person's eye color, it turns out that 1% of people in the world have green eyes ("What percentage of," 2013 ). Consider a group of 20 people.

1. Find the probability that none have green eyes, and find the probability that nine have green eyes.
2. Find the probability that at most three have green eyes.
3. Find the probability that more than 6 people have green eyes.
4. Find the probability that at least four have green eyes.
5. In Europe, four people out of twenty have green eyes. Is this unusual? What does that tell you?

*Answer 1 available at the end of this section.*

### Technology: To find the Binomial Probability

We can use Technology such as a TI calculator, Excel, StatCrunch, R, etc., to find the probability rather than using a formula. In this section, you will find **GeoGebra Stat Calculator** and **Online Statistics Calculator** that can be used to find binomial probabilities. Watch the attached video at the bottom of the page to learn how to find probability using the binomial distribution.

Note the following:

$$P(x < a) = P(x \leq a - 1)$$

$$P(x > a) = P(x \geq a + 1)$$

For example,

$$P(x < 10) = P(x \leq 9)$$

$$P(x > 15) = P(x \geq 16)$$

### Your Turn 5.2.4: Find Binomial Probability

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Online Statistic Calculator

[Online Statistics Calculator](#)

[GeoGebra Statistics Calculator](#)

### Example 5.2.5: Find the Binomial Probability

It has been stated that about 41 of adult workers have a high school diploma but do not pursue any further education. If 10 adult workers are randomly selected, construct a binomial distribution of the number of workers who have a high school diploma but do not pursue any further education. Find the probability that at most 6 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 not pursue any further education?

Answer available at the end of this section.

### Your Turn 5.2.5: Create Binomial Distribution Table

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Exercise 5.2.6: Find the Probability Using Binomial

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.

1. Why is the probability distribution for  $X$  binomial?
2. Find the probability that 25 adults in the sample prefer saving over spending
3. Find the probability that at most 20 adults prefer saving
4. Find the probability that more than 30 adults prefer saving

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

Answer 3 available at the end of this section.

Answer 4 available at the end of this section.

### Your Turn 5.2.6: Find the Probability Using Binomial

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Mean Standard Deviation of the Binomial Probability Distribution

### Formula to Find the Mean and Standard Deviation

The outcomes of a binomial experiment fit a **binomial probability distribution**. Suppose  $X$  is a random variable and  $x$  is the number of successes obtained in the  $n$  independent trials.  $p$  is the probability of success and  $1 - p$  is the probability of failure.

The mean,  $\mu_{\bar{x}}$  of the binomial probability distribution is

$$\mu_{\bar{x}} = np$$

The standard deviation,  $\sigma_{\bar{x}}$ , is then

$$\sigma_{\bar{x}} = \sqrt{np(1-p)}$$

If you list all possible values of  $x$  in a Binomial distribution, you get the **Binomial Probability Distribution**. You can draw a histogram of the binomial probability distribution and find its mean, variance, and standard deviation.

## Unusual Observation and Empirical Rule

For a fixed  $p$ , as the number of trials  $n$  in a binomial experiment increases, the probability distribution of the random variable  $X$  becomes bell-shaped and the number of trials increases. The two conditions should be satisfied for  $n$  to be approximately normally distributed in a binomial distribution.

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

### Empirical Rule: Unusual Observation

An unusual observation in a binomial distribution is an outcome with a very low probability. As a rule of thumb, if two of the following conditions satisfy the interval  $[\mu_{\bar{x}} - 2\sigma_{\bar{x}}, \mu_{\bar{x}} + 2\sigma_{\bar{x}}]$  is called a usual interval.

$$np \geq 10$$

$$n(1-p) \geq 10$$

The minimum usual value

$$\mu_{\bar{x}} - 2\sigma_{\bar{x}}$$

The maximum usual value

$$\mu_{\bar{x}} + 2\sigma_{\bar{x}}$$

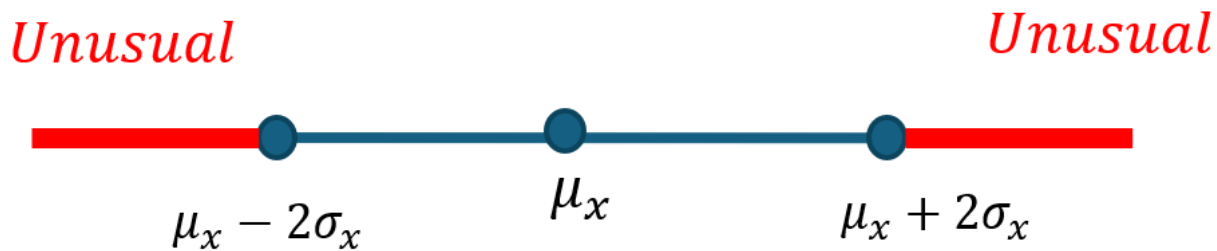


Figure 5.2.2: Usual Interval

### Example 5.2.7: Probability Using Binomial

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  $X$  = be the number of shots that scored points, and it is binomial.

1. Using the formulas, calculate the mean and standard deviation of  $X$ .

2. Verify that the binomial distribution is approximately normal, and draw a histogram to see whether the distribution is symmetric.
3. Use the rule of thumb to find the usual interval.
4. Would it be unusual if DeAndre scored 75 or 40 out of 80 shoots?

Answer 1 available at the end of this section.

### Your Turn 5.2.7: Check Unusual Observations

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## References

1. "Access to electricity (% of population)," The World Bank, 2013. Available online at [Word Bank Data](http://data.worldbank.org) [data.worldbank.org] (accessed May 15, 2015).
2. "Distance Education." Wikipedia. Available online at [Wikipedia Link](http://en.Wikipedia.org) [en.Wikipedia.org] (accessed May 15, 2013).
3. "NBA Statistics – 2013," ESPN NBA, 2013. Available online at [Gallup dot com](http://espn.go.com) [espn.go.com] (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 [Gallup dot com](http://www.gallup.com) [www.gallup.com] (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 [UCLA](http://heri.ucla.edu) [heri.ucla.edu] (accessed May 15, 2013).
6. "The World FactBook," Central Intelligence Agency. Available online at [UCLA EDU](http://www.cia.gov) [www.cia.gov] (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](http://www.cancer.org/cancer/pancrea...key-statistics) (accessed May 15, 2013).

## Answers

### Example 5.2.1: Check if the Experiment is Binomial or Not?

Random Variable X: number of heads.

Two outcomes: head or tail

Independent trials,  $n = 10$  fixed number of trials.

Probability of success (head) is fixed in each trial, which is  $\frac{1}{2}$ .

So, this is a **binomial experiment**, and variable X is a binomial random variable.

The value of  $x$  are:  $x = 0, 1, \dots, 10$ .

Random Variable X: number of correct answers.

Either you get the question right, or you get it wrong, so there are only two outcomes.

In this case, the success is getting the question right. Getting the first question right has no effect on getting the second or third question right, the trials are independent.

There are 15 questions, and each is a trial, so the total number of trials is fixed ( $n=15$ ).

In this case,  $p = \frac{1}{4}$ . So, this is a **binomial experiment**.

The value of  $x$  are:  $x = 0, 1, \dots, 15$ .

Random Variable X: number of hearts.

Two outcomes: getting an ace or not.

But not an independent trial, since drawing one card without replacement affects the probability of the second draw being a heart.

This is not a binomial experiment since trials are dependent and  $p$  is not fixed in each trial.

The value of  $x$  are:  $x = 0, 1, \dots, 5$ .

Random Variable X: number of correct answers. Either you get the question right, or you get it wrong, so there are only two outcomes.

In this case, the success is getting the question right. Getting the first question right has no effect on getting the second or third question right; the trials are independent.

There are 10 questions, each a trial, so there is a fixed number of trials.

But  $p$  is not fixed in each trial since  $p = \frac{1}{4}$  (For multiple choice question) and  $p = \frac{1}{2}$  (For T/F question). So, this is **not a binomial experiment**.

The value of  $x$  are:  $x = 0, 1, \dots, 10$ .

#### Example 5.2.2: Binomial Experiment Terminology

Random Variable  $X$ : the number of statistics students who do their homework on time.

Either the student completes homework on time or not, so there are only two outcomes.

Finding one student who does the homework on time does not depend on the other students; the trials are independent.

There are 50 students, so the total number of trials is fixed. In this trial, the probability of a success is  $p = 0.70$ .

So, this is a **binomial experiment**.

#### Example 5.2.2: Binomial Experiment Terminology

Out of 50, either 0 student do homework on time or 1 students do homework on time, ..., all 50 students do homework on time. So  $x = 0, 1, 2, \dots, 50$

#### Example 5.2.2: Binomial Experiment Terminology

does not do the homework on time

#### Example 5.2.2: Binomial Experiment Terminology

At least means more than or equal. So in a probability question, at least is translated as

$$P(x \geq 40)$$

#### Example 5.2.3: Calculating Binomial Probabilities (Using Formula)

Random variable:  $X$  = number of people with green eyes, and it is binomial

Success here is people with green eyes and the probability of success,  $p = 0.01$ . The probability that none have green eyes is translated to

$P(x = 0)$ . So using the formula

$$P(x = 0) = {}_{20}C_0(0.01)^0(0.99)^{20-0} \approx 0.8179$$

You can find  $P(x = 0)$  using an online statistics calculator. Here are the steps.

1. Put the cursor under distribution and click "Binomial"
2. Type  $n = 20$  and  $p = 0.01$  and check  $P(x = a)$  box.
3. Type  $P(x = 0)$ , the number in the bottom right box above is the desired probability.

$$P(x = 0) = 0.8179$$

The probability that nine has green eyes is translated to

$P(x = 9)$ . So using the formula

$$P(x = 9) = {}_{20}C_9(0.01)^9(0.99)^{20-9} \approx 1.50 \times 10^{-13} \approx 0.000$$

Or you can use an online statistics calculator as shown above.

At most three means that three is the highest value you will have. Find the probability of  $x$  is less than or equal to three.

$$\begin{aligned} P(x \leq 3) &= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) \\ &= {}_{20}C_0(0.01)^0(0.99)^{20} + {}_{20}C_1(0.01)^1(0.99)^{19} + {}_{20}C_2(0.01)^2(0.99)^{18} + {}_{20}C_3(0.01)^3(0.99)^{17} \\ &\approx 0.818 + 0.165 + 0.016 + 0.001 > 0.999 \end{aligned}$$

The reason the answer is written as being greater than 0.9999573 is that the answer is actually 0.9999573791, and when that is rounded to three decimal places, you get 1. But 1 means that the event will happen, when in reality, there is a slight chance that it won't happen. It is best to write the answer as greater than 0.9999573 to represent that the number is very close to 1, but isn't 1.

You can find  $P(x \leq 3)$  using an online statistics calculator. Here are the steps.

1. Put the cursor under distribution and click "Binomial"
2. Type  $n = 20$  and  $p = 0.01$  and check  $P(x \leq b)$  box.
3. Type  $P(X \leq 3)$ , the number in the bottom right box is the desired probability.

$$P(x \leq 3) = 0.9999573$$

Find  $P(x > 6)$ . Note that  $P(x > 6)$  is same as  $P(x \geq 7)$

1. Put the cursor under distribution and click "Binomial"
2. Type  $n = 20$  and  $p = 0.01$  and check  $P(x \geq a)$  box.
3. Type  $P(7 \geq X)$ , the number in the bottom right box is the desired probability.

$$P(x > 6) = 6.999...e^{-10} \approx 0.0000000007 \approx 0$$

At least four means four or more. Find the probability of  $x$  being greater than or equal to four.

$$\begin{aligned} P(x \geq 4) \\ P(x \geq 4) = 0.001 \end{aligned}$$

1. Put the cursor under distribution and click "Binomial"
2. Type  $n = 20$  and  $p = 0.01$  and check  $P(x \geq a)$  box.
3. Type  $P(3 \geq X)$ , the number in the bottom right box is the desired probability.

$$P(x \leq 3) = 0.0010$$

Since the probability of finding four or more people with green eyes is much less than 0.05, it is unusual to find four people out of twenty with green eyes. That should make you wonder if the proportion of people in Europe with green eyes is more than the 1 for the general population. If this is true, then you may want to ask why Europeans have a higher proportion of green-eyed people. That, of course, could lead to more questions.

### Example 5.2.5: Find the Binomial Probability

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, ..., 10 where  $n = 10, p = 0.41$ . Using an online GeoGebra statistics calculator, we can construct the following binomial distribution of  $X$ .

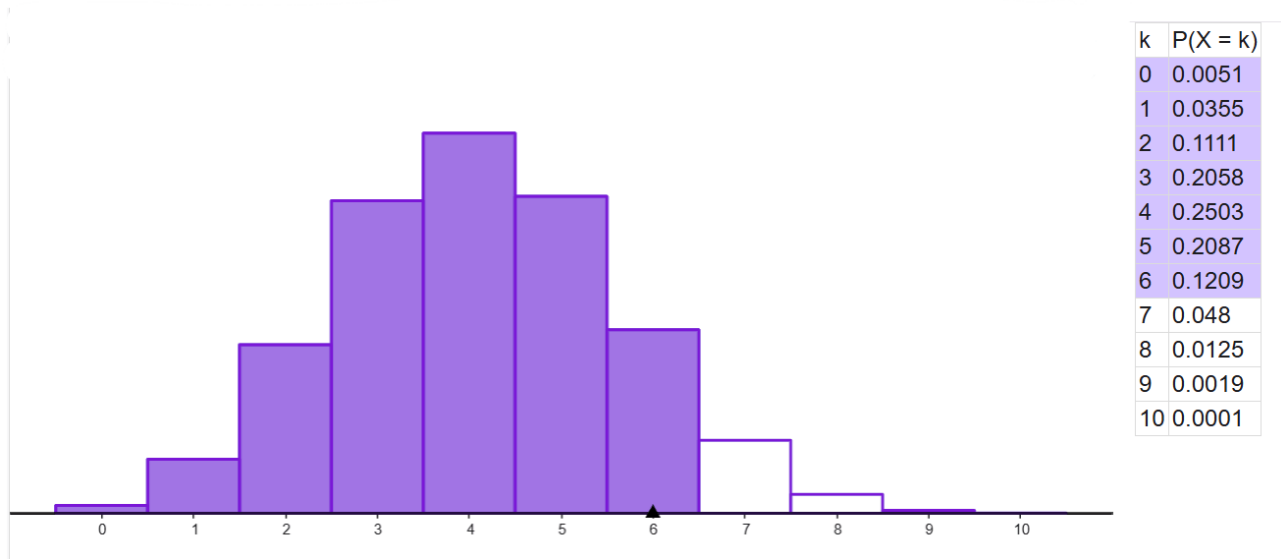


Figure 5.2.1: Binomial Distribution

At most 6 is translated as  $P(x \leq 6)$

$$P(x \leq 6) = 0.0051 + 0.0355 + 0.1111 + 0.2058 + 0.2503 + 0.2087 + 0.1209 = 0.9374$$

Or you can use statistics online calculator to find

$$P(x \leq 6) = 0.9374$$

### Exercise 5.2.6: Find the Probability Using Binomial

Random Variable  $X$  = The number of American adults who prefer saving to spending

Two outcomes: saving or spending

Independent trials,  $n = 50$  fixed number of trials.

Probability of success (saving) is fixed in each trial, which is 0.60.

So, this is a **binomial experiment**, with  $X$  a binomial random variable.

The value of  $x$  are:  $x = 0, 1, \dots, 50$ .

### Exercise 5.2.6: Find the Probability Using Binomial

The probability that 25 adults prefer saving over spending is translated as  $P(x = 25)$ .

Here are the steps to find the probability in an online statistics calculator.

1. Put the cursor under distribution and click "Binomial"
2. Type  $n = 50$  and  $p = 0.60$  and check  $P(x = a)$  box.
3. Type  $P(x = 25)$ , the number in the bottom right box is the desired probability.

$$P(x = 25) = 0.0405$$

### Exercise 5.2.6: Find the Probability Using Binomial

The probability that at most 20 adults prefer saving over spending is translated as  $P(x \leq 20)$ .

Here are the steps to find the probability in an online statistics calculator.

1. Put the cursor under distribution and click "Binomial"
2. Type  $n = 50$  and  $p = 0.60$  and check  $P(x \leq b)$  box.
3. Type  $P(x \leq 20)$ , the number in the bottom right box is the desired probability.

$$P(x \leq 20) = 0.0034$$

### Exercise 5.2.6: Find the Probability Using Binomial

The probability that at most 20 adults prefer saving over spending is translated as  $P(x > 30) = P(x \geq 31)$ .

Here are the steps to find the probability in an online statistics calculator.

1. Put the cursor under distribution and click "Binomial"
2. Type  $n = 50$  and  $p = 0.60$  and check  $P(x \geq a)$  box.
3. Type  $P(x \geq 31)$ , the number in the bottom right box is the desired probability.

$$P(x > 30) = 0.4465$$

### Example 5.2.7: Probability Using Binomial

Mean and standard deviation of binomial random variable  $X$  is give as below

$$\begin{aligned}\text{Mean: } \mu_{\bar{x}} &= np \\ &= 80(0.613) \\ &= 49.04 \\ \text{Standard Deviation: } \sigma_{\bar{x}} &= \sqrt{np(1-p)} \\ &= \sqrt{80(0.613)(1-0.613)} \\ &\approx 4.36\end{aligned}$$

Let's check the conditions:

$$\begin{aligned}np &= 80(0.613) &= 49.04 \geq 10 \\ n(1-p) &= 80(1-0.613) &= 30.96 \geq 10\end{aligned}$$

So both criteria are verified, and the probability histogram is given below. And the binomial distribution is symmetric.

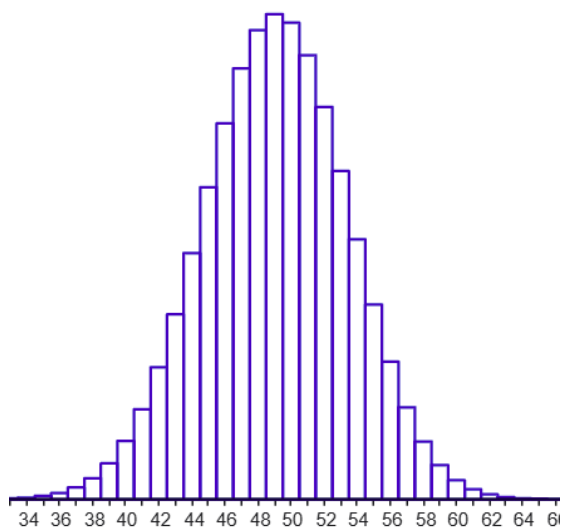


Figure 5.2.3: Probability Histogram

Minimum usual value

$$\mu_{\bar{x}} - 2\sigma_{\bar{x}} = 49.04 - 2(4.36) = 31.37$$

Maximum usual value

$$\mu_{\bar{x}} + 2\sigma_{\bar{x}} = 49.04 + 2(4.36) = 57.76$$

So usual interval is

$$[31.37, 57.76]$$

The observation 75 is unusual since it is outside the usual interval, and 40 is not unusual.

This page titled [5.2: Binomial Probability Distributions](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Kathryn Kozak](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10: Exponential and Logarithmic Functions](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.
- [CustomerActivityOverview](#) has no license indicated.
- [14.4: The Development and Use of Different Number Bases](#) by [David Lippman](#) is licensed [CC BY-SA 3.0](#). Original source: <http://www.opentextbookstore.com/mathinsociety>.
- [11.1: Prelude to Conics](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/intermediate-algebra-2e>.

## 6: 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 ( $\mu$ ) and the standard deviation ( $\sigma$ ).

6.1: Properties of the Normal Distribution

6.2: Normal Distribution and Applications

### Contributors

- [Template:ContribOpenStax](#)

---

This page titled [6: 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.

## 6.1: Properties of the Normal Distribution

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal bell curve is ubiquitous throughout statistics. Indeed, it is so common that people often know it as the **normal curve** or **normal distribution**, shown in Figure 6.1.1. It is also known as the **Gaussian distribution** after Frederic Gauss, the first person to formalize its mathematical expression. Variables such as SAT scores and the heights of US adult males closely follow the normal distribution.

Many variables are nearly normal, but none are exactly normal. In applied statistics, describing real data as **approximately normal** is usually the more precise and accurate phrasing. In this text, whenever "normal" appears, it means approximately normal.

### Normal and Standard Normal Distribution Model

The normal distribution model always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation. As you can probably guess, changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve.

#### Side-by-Side Comparison Between Normal RV and Standard Normal RV

Side-by-Side Comparison Between Normal RV and Standard Normal RV

Table 6.1.1: Comparison Between Normal RV and Standard Normal RV

| Feature                            | Normal Random Variable (X)            | Standard Normal Random Variable (Z) |
|------------------------------------|---------------------------------------|-------------------------------------|
| <b>Mean</b> $\mu$                  | Any real number (e.g., 100, -5, 72.5) | Exactly 0                           |
| <b>Variance</b> $\sigma^2$         | Any positive number                   | Exactly 1                           |
| <b>Standard Deviation</b> $\sigma$ | Any positive number (e.g., 15, 2.1)   | Exactly 1                           |
| <b>Center of the Bell Curve</b>    | Centered at $\mu$                     | Centered exactly at 0               |

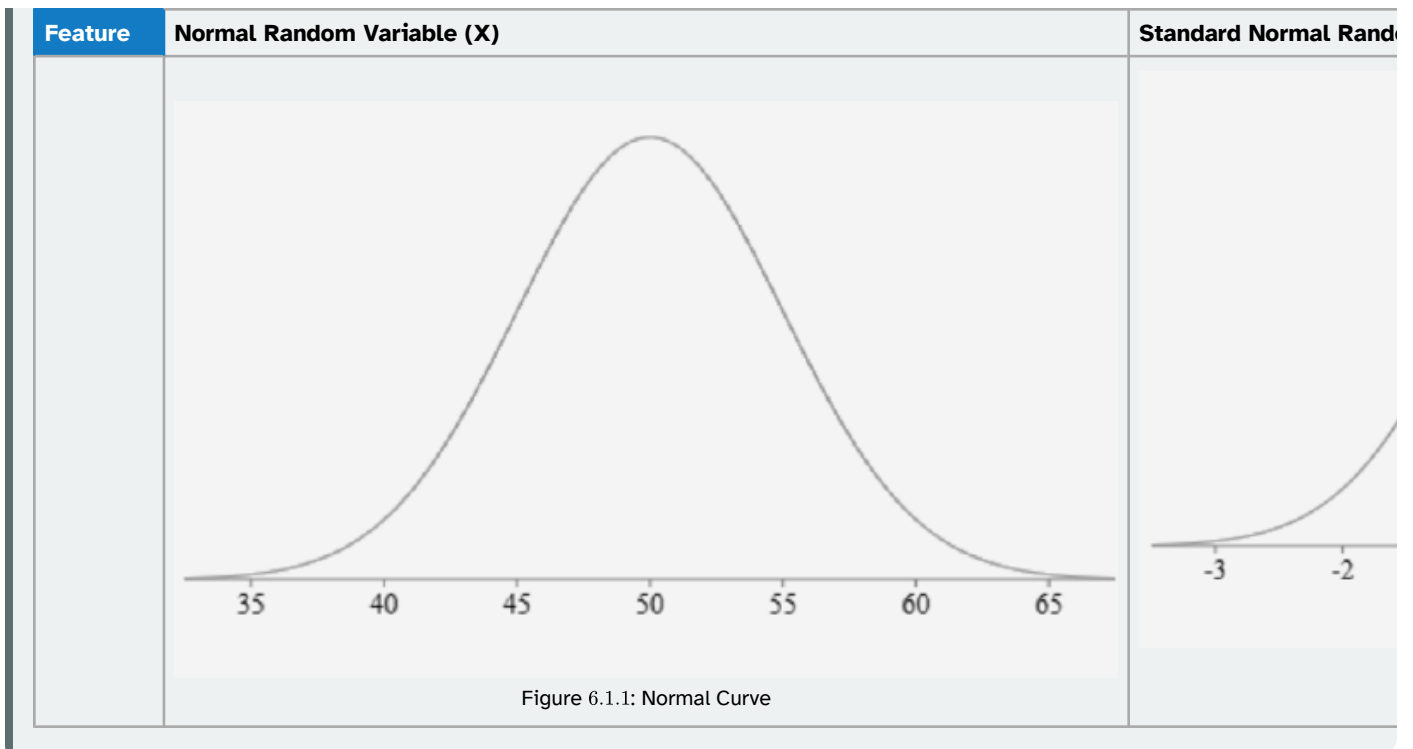


Figure 6.1.2 shows the normal distribution with mean 0 and standard deviation 1 in the figure 6.1.1 shows the normal distributions with mean 50 and standard deviation 5 in the right panel.

The following figure 6.1.3 shows these distributions on the same axis.

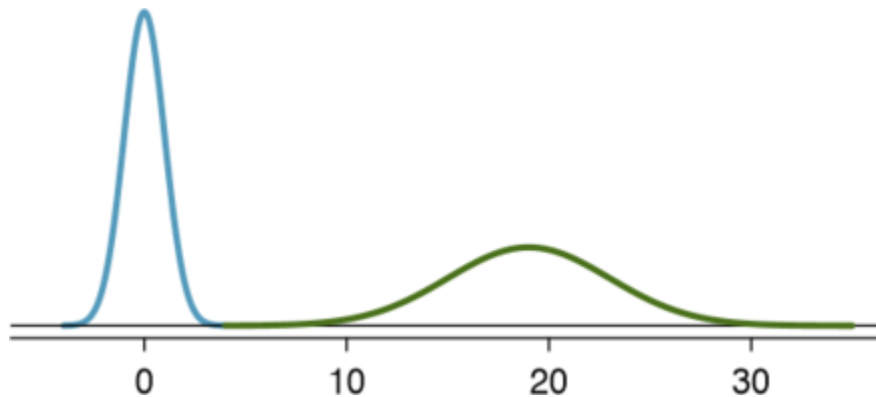


Figure 6.1.3: But plotted together and on the same scale.

If a normal distribution has mean  $\mu$  and standard deviation  $\sigma$ , we may write the distribution as Normal with  $(\mu, \sigma)$ . The two distributions in figure 6.1.2 and figure 6.1.1 can be written as

Normal with  $(\mu = 0, \sigma = 1)$

and

Normal with  $(\mu = 50, \sigma = 5)$

Because the mean and standard deviation fully describe a normal distribution, they are called the **distribution's parameters**.

Actually, we learn in Chapter 2 that we can convert data  $x$  to its  $z$  score using the following formula. This process of converting  $x$  to a standard value is called standardizing  $x$ .

## Standardizing with Z Scores

The *Z score* of an observation *Z* is defined as the number of standard deviations it falls above or below the mean. If the observation is one standard deviation above the mean, its *Z* score is 1. If it is 1.5 standard deviations below the mean, then its *Z* score is  $-1.5$ .

### Definition: The Z score

The z-score of an observation is the number of standard deviations it is above or below the mean. We compute the z score for an observation  $x$  that follows a distribution with mean  $\mu$  and standard deviation  $\sigma$  using

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

Also, we can find the  $x$  value given the z-score, mean, and standard deviation.

$$x = \mu + \sigma z$$

Observations above the mean always have positive *Z* scores, while those below the mean have negative *Z* scores. If an observation is equal to the mean.

### Example 6.1.1: Find and Interpret z-score

Let *X* represent a random variable from Normal with  $\mu = 3$  and  $\sigma = 2$ , and suppose we observe  $x = 5.19$ .

- Find the z-score of  $x$  and interpret it.
- Find the data value whose z-score is  $-1.27$ .

*Answer a available at the end of this section.*

### Your Turn 6.1.1: Find and Interpret z-score

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Probability, Proportion, and Area

### Example of Normal Data

The following data set contains statistics on the final exam scores at the university, with 170 students. The scores are normally distributed with mean  $\mu = 74.7$  and standard deviation  $\sigma = 17.1$ .

Table 6.1.2: Final Exam Score of 170 students

| Col 1 | Col 2 | Col 3 | Col 4 | Col 5 | Col 6 | Col 7 | Col 8 | Col 9 | Col 10 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| 63.2  | 63.4  | 63.6  | 63.8  | 64.1  | 64.3  | 64.5  | 64.7  | 64.9  | 64.9   |
| 100.9 | 101.8 | 102.8 | 103.9 | 105.1 | 106.5 | 64.9  | 110.2 | 113.1 | 113.1  |

| Col 1 | Col 2 | Col 3 | Col 4 | Col 5 | Col 6 | Col 7 | Col 8 | Col 9 | Col 10 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| 94.6  | 52.5  | 65    | 66    | 56.1  | 56.2  | 97.9  | 98.6  | 99.3  | 99.3   |
| 90.7  | 91.1  | 91.4  | 91.7  | 92.1  | 58.2  | 92.9  | 93.3  | 93.7  | 93.7   |
| 87.9  | 88.2  | 88.4  | 88.7  | 53.5  | 89.2  | 89.5  | 89.8  | 90.1  | 90.1   |
| 85.5  | 95    | 86    | 53    | 52    | 86.7  | 86.9  | 87.1  | 87.4  | 87.4   |
| 83.3  | 36    | 83.7  | 83.9  | 84.1  | 84.3  | 84.6  | 84.9  | 85.1  | 85.1   |
| 81.3  | 81.5  | 81.7  | 54.6  | 82.1  | 82.3  | 82.5  | 82.7  | 82.9  | 82.9   |
| 79.3  | 35    | 45    | 79.9  | 80.1  | 80.3  | 28    | 80.7  | 80.9  | 80.9   |
| 75.3  | 25    | 75.7  | 75.9  | 54.1  | 76.3  | 43    | 76.7  | 76.9  | 76.9   |
| 73.3  | 73.5  | 73.7  | 122   | 74.1  | 74.3  | 74.5  | 74.7  | 74.9  | 74.9   |
| 71.3  | 71.5  | 71.7  | 71.9  | 72.1  | 72.3  | 72.5  | 72.7  | 72.9  | 72.9   |
| 69.3  | 69.5  | 69.7  | 69.9  | 70.1  | 70.3  | 70.5  | 70.7  | 70.9  | 70.9   |
| 67.3  | 67.5  | 38    | 67.9  | 68.1  | 43    | 68.5  | 68.7  | 68.9  | 68.9   |
| 65.3  | 65.5  | 65.7  | 65.9  | 66.1  | 66.3  | 66.5  | 66.7  | 66.9  | 66.9   |
| 52.2  | 49.2  | 58.4  | 45.6  | 55.5  | 50.9  | 56.2  | 76.7  | 44.9  | 44.9   |
| 42    | 53    | 77.7  | 124   | 58    | 78.3  | 78.5  | 78.7  | 78.9  | 78.9   |

The frequency and relative frequency table is given below.

Table 6.1.3: Frequency and Relative Frequency Distribution Table

| Class     | Frequency | Relative Frequency (proportion) |
|-----------|-----------|---------------------------------|
| 20 – 30   | 2         | 0.012                           |
| 30 – 40   | 3         | 0.018                           |
| 40 – 50   | 8         | 0.047                           |
| 50 – 60   | 16        | 0.094                           |
| 60 – 70   | 35        | 0.206                           |
| 70 – 80   | 41        | 0.241                           |
| 80 – 90   | 37        | 0.218                           |
| 90 – 100  | 17        | 0.10                            |
| 100 – 110 | 6         | 0.035                           |
| 110 – 120 | 3         | 0.017                           |
| 120 – 130 | 2         | 0.012                           |

The following graph shows a histogram of the data above.

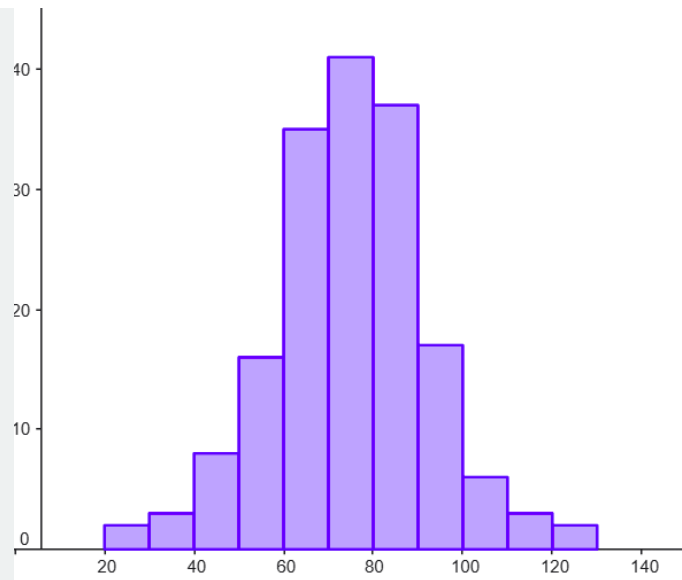


Figure 6.1.4: Normal Distribution

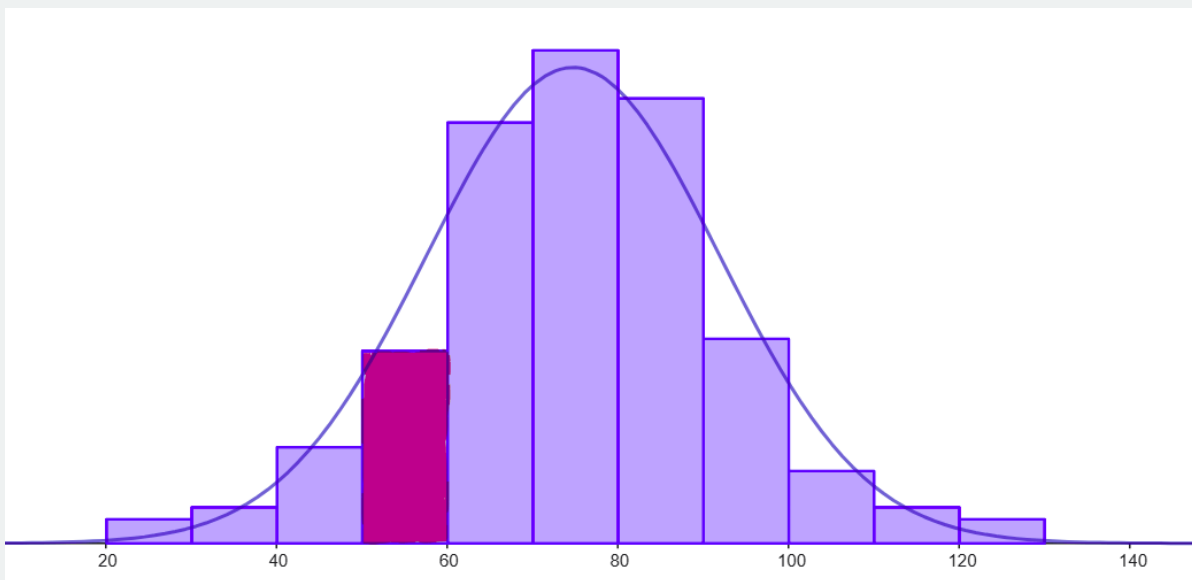


Figure 6.1.5: Normal Density Curve

### Probability, Proportion, and Area

In a normal distribution, **probability**, **proportion**, and **area** are three different ways of expressing the exact same concept. Because the total area under the smooth, bell-shaped curve is exactly equal to 1 (or 100%), any specific region under the curve represents a fraction of the whole. For instance, if you calculate the shaded area under the curve between two specific values, that physical space represents the proportion of the population falling within that range.

## 68-95-99.7 Rule and Percentile

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when making a quick estimate without a

calculator or Z table.

# Empirical Rule, Area and Proportion

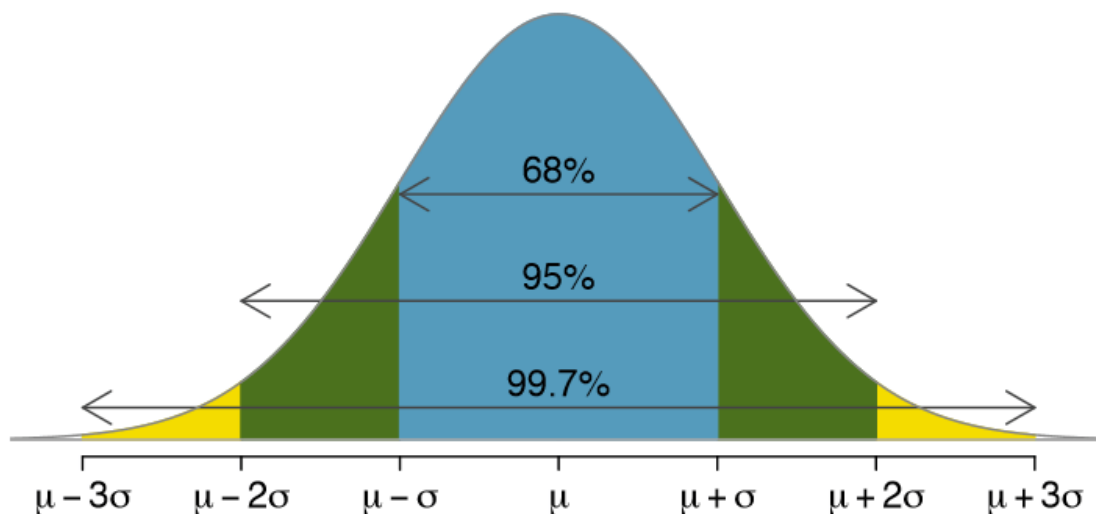


Figure 6.1.6: Empirical Rule

You are likely thinking of the **Empirical Rule** (also known as the 68–95–99.7. For a standard normal distribution with a **mean**,  $\mu = 0$  and a **standard deviation**,  $\sigma = 1$ , the  $x$  values correspond directly to **z scores**  $-3, -2, -1, 0, 1, 3$ .

## STANDARD NORMAL CURVE: PROPORTIONS BY REGION

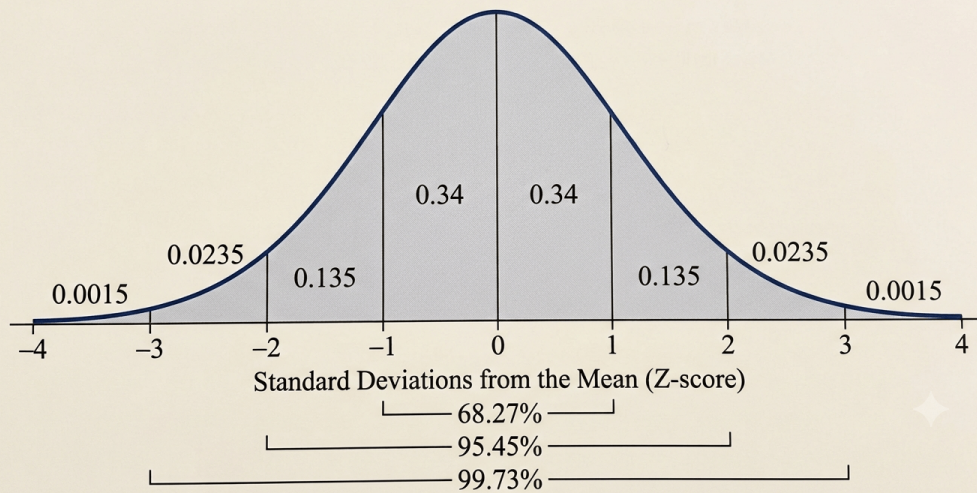


Figure 6.1.7: Area by Proportion

### Your Turn 6.1.2: Shade the Area

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Properties of the Normal Curve

- Shape and Symmetry Bell-Shaped:** The curve rises to a single peak in the center and tapers off smoothly on both sides. So it is symmetric about the mean  $\mu$ .
- The "Three M's" Coincide:** In a perfect normal distribution, the mean, median, and mode are all equal and located at the exact center of the curve.
- Inflection Points:** The curve changes its direction of curvature (concavity) at exactly one standard deviation above and below the mean  $\mu - 1\sigma$  and  $\mu + 1\sigma$ . This is where the curve shifts from being "cupped downward" at the peak to "cupped upward" at the tails
- Total Area and Probability: Total Area = 1:** The total area under the normal curve represents 100% of the probabilities for the dataset. Therefore, the area under the curve is exactly equal to 1.
- $\frac{1}{2}$  Split:** Because it is perfectly symmetric, exactly 50% of the area lies to the left of the mean, and 50% lies to the right. Or area under the curve on each side of  $\mu$  equal to  $\frac{1}{2}$ .
- Asymptotic to the Horizontal Axis:** The tails of the curve flatten out and extend infinitely to the left and right, getting closer and closer to the horizontal axis but never actually touching or crossing it. No matter how far you go, the probability never reaches absolute zero.
- The Empirical Rule (The 68 – 95 – 99.7 Rule):** For any normal distribution

About 68% of the data falls within 1 standard deviation of the mean.

About 95% of the data falls within 2 standard deviations of the mean.

About 99.7% of the data falls within 3 standard deviations of the mean.

### Percentile and Interpretation of the Empirical Rule

In the above example of 170 statistics score. Look at table 6.1.1.

Since

$$\mu - 1\sigma = 74.7 - 17.1 = 57.6 \quad \text{and} \quad \mu + 1\sigma = 74.7 + 17.1 = 91.8$$

68% of the final exam score falls between 57.6 and 91.8.

Since

$$\mu - 2\sigma = 74.7 - 2(17.1) = 40.5 \quad \text{and} \quad \mu + 2\sigma = 74.7 + 2(17.1) = 108.9$$

95% of the final exam score falls between 40.5 and 108.9.

Since

$$\mu - 3\sigma = 74.7 - 3(17.1) = 23.4 \quad \text{and} \quad \mu + 3\sigma = 74.7 + 3(17.1) = 126$$

99.7% of the final exam score falls between 23.4 and 126.

1. 50% data are less than or equal to 74.70, we call 74.70 is 50<sup>th</sup> percentile. Or  $P_{50} = 74.70$
2. 16% data are less than or equal to 57.60, we call 57.60 is 16<sup>th</sup> percentile. Or  $P_{16} = 57.60$

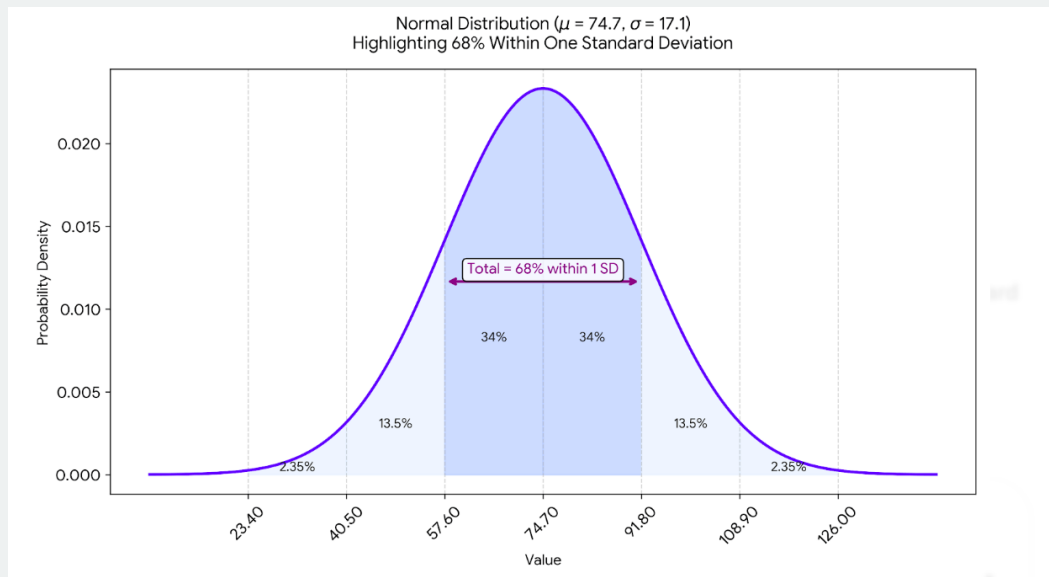


Figure 6.1.8: Empirical Rule

### Example 6.1.3: Understanding Normal Curve

The set of data in table 6.1.2 consists of statistics on final exam scores at the university, with 170 students. Their scores are normally distributed with mean  $\mu = 74.7$  and standard deviation  $\sigma = 17.1$ . Refer to the figure 6.1.8 and table 6.1.3 to answer the following question.

1. Find the inflection point for the normal curve. What percentage of the student's score is between the two inflection points?
2. What proportion of test-takers score between 50 – 60?
3. The probability that a randomly selected student's score is between 50 – 60?
4. Find the area under the normal curve to the right of 100.
5. Find  $P(X < 40)$
6. What is 84<sup>th</sup> percentile?
7. Find the area under the normal curve to the left of 40.50.
8. If the area under the curve to the right of 108.90 is 0.025. What are the two interpretations of this result?

Answer 1 available at the end of this section.

### Your Turn 6.1.3: Understanding Normal Curve

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Area to the Right, Left, or in Between

In a normal curve, the area to the left of a  $x$ -value represents the probability that a value is less than that  $x$ -score, while the area to the right represents the probability that a value is greater than that  $x$ -score. The area between two  $x$ -scores represents the probability that a value lies between those two values. Since the total area under the normal curve is 1, all probabilities are found by calculating areas under the curve. There are three situations involving the area.

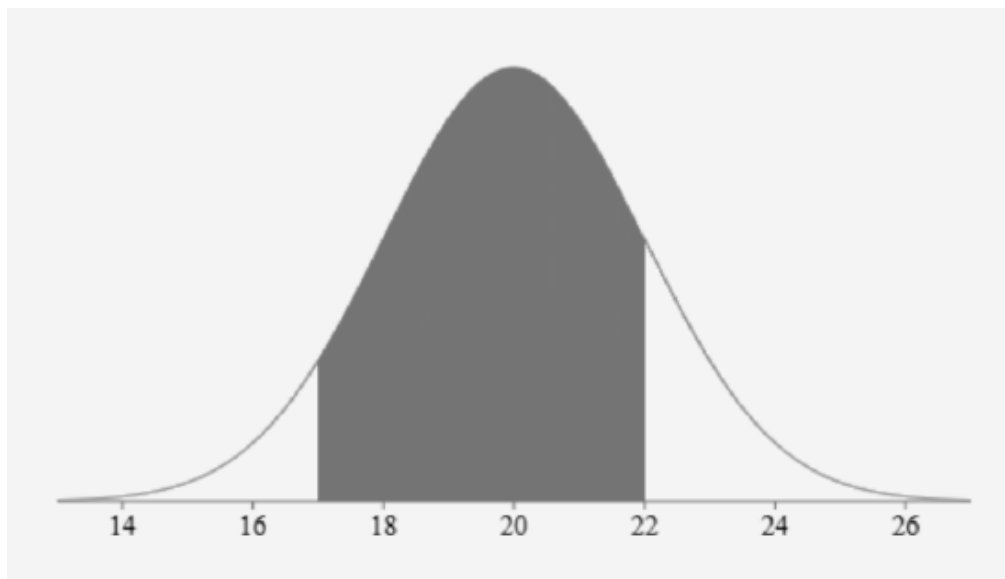


Figure 6.1.9:  $P(17 \leq x \leq 22)$

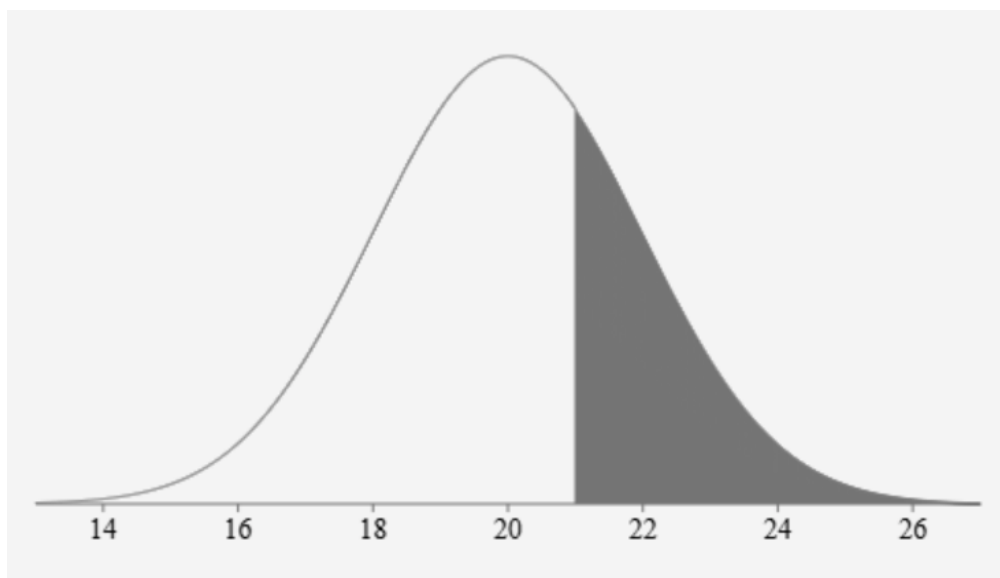


Figure 6.1.10:  $P(x > 21)$

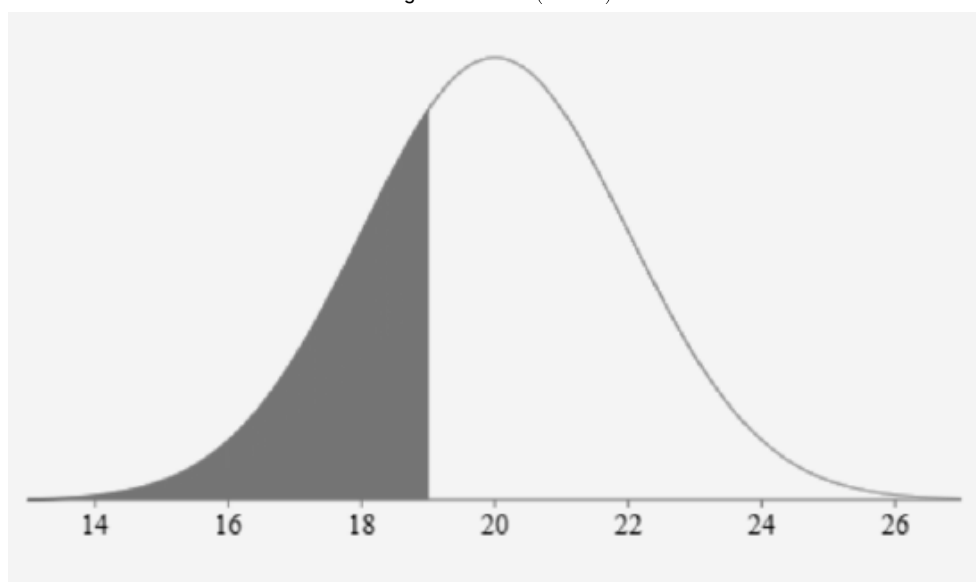


Figure 6.1.11:  $P(x < 17)$

The above three figures show the three different types of area under the normal curve with a mean 20 and a standard deviation 2.

### Remember

In a **normal distribution**,

$P(x > a) = P(x \geq a)$  or  $P(x < a) = P(x \leq a)$  because a normal distribution is **continuous**.

$P(x \leq)$

**TIP: Always draw a Diagram First with labeled**

For any normal probability problem, always draw and label the normal curve (or standard normal curve), including the mean and any relevant values, and shade the area of interest first. This helps you identify which probability or area you are finding and provides a visual estimate of the answer.

**Example 6.1.4: Shade the Area**

Edward earned a 1400 on his SAT. 36.9% of other scores are lower than his. Shade the region that corresponds to the following statement.

$$P(x \leq 1400) = 0.369$$

What is the Eduardo score percentile?

Answer available at the end of this section.

**Your Turn 6.1.4: Shade the Area**

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Answers

**Example 6.1.1: Find and Interpret z-score**

Its Z score is given by

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{5.19 - 3}{2} \\ &= \frac{2.19}{2} = 1.095 \end{aligned}$$

$$z = \frac{x - \mu}{\sigma} = \frac{5.19 - 3}{2} = \frac{2.19}{2} = 1.095.$$

Observation  $x$  is 1.095 standard deviations above the mean. We know it must be above the mean since  $z$  is positive.

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ -1.27 &= \frac{x - 3}{2} \\ -1.27 \times 2 &= \frac{x - 3}{2} \times 2 \\ -2.54 &= x - 3 \\ 0.46 &= x \end{aligned}$$

**Example 6.1.3: Understanding Normal Curve**

Inflection points are the points that are 1 standard deviation above or below the mean.

So 57.60 and 91.80 are inflection points. 68%: Figure 6.1.8

0.094: Table 6.1.3

$P(50 < x < 60) = 0.094$ : Table 6.1.3

$0.035 + 0.017 + 0.012 = 0.064$ : Table 6.1.3

$P(40 < x < 50) = 0.047$ : Table 6.1.3

84% score are less than or equal to 91.80, so 84<sup>th</sup> percentile is 91.80: Figure 6.1.8

$0.15\% + 2.35\% = 2.5\%$ : Empirical Rule: Figure 6.1.8

2.5% score are more than 108.90 or  $P(x > 108.90) = 0.025 = 2.5\%$

Interpretation 1: If one student is selected randomly, the probability that the student scores more than 108.9 is 0.025.

Interpretation 2: The proportion of students who score more than 108.9 is 0.025.

#### Example 6.1.4: Shade the Area

First, a picture is needed. Edward's percentile is the proportion of people who score below 1400. This proportion is 0.0369  $\approx$  0.37. In other words, 37% are the scores that are less than or equal to 1400. Please refer to the figure 6.1.12.

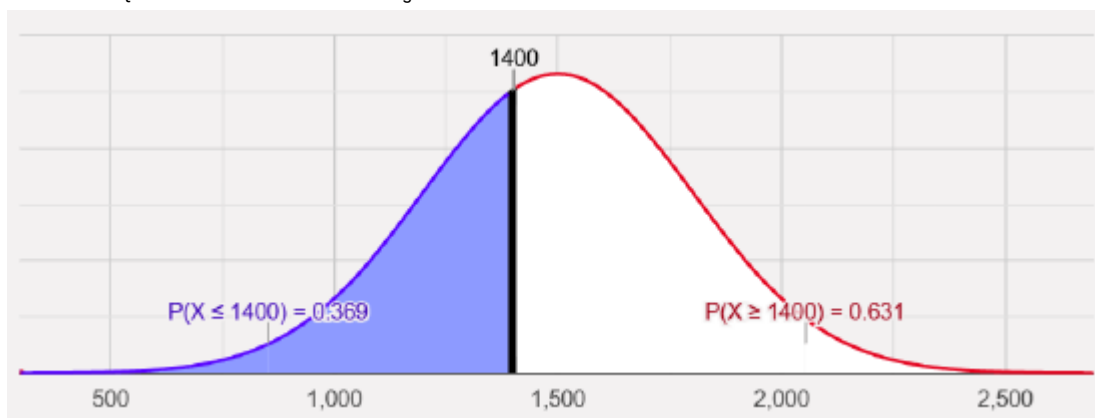


Figure 6.1.12:  $P(x \leq 1400)$

Edward's score is at the 37<sup>th</sup> percentile.

This page titled [6.1: Properties of the Normal Distribution](#) is shared under a [CC BY-SA 3.0](#) license and was authored, remixed, and/or curated by [David Diez](#), [Christopher Barr](#), and [Mine Çetinkaya-Rundel](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.7: Invertibility](#) by [Isaiah Lankham](#), [Bruno Nachtergaele](#), and [Anne Schilling](#) has no license indicated.

## 6.2: Normal Distribution and Applications

### Recall from the Previous Section

#### Recall from the Previous Section

Recall from the previous section: **area = proportion = probability.**

For example, in the following normal density curve, the number  $(0.69)$  can be represented in three different ways

1.  $P(X \geq 95) = 0.69$
2. The proportion of data values that are more than  $(95)$  is  $(0.69)$ . Or  $(69\%)$  data values are more than  $(95)$ .
3. Area under the curve to the right of  $(95)$  is  $(0.69)$ . In other words, the area on the left of  $(95)$  is  $(0.31)$ . In this case we call  $(x=95)$  is  $(31)$ th percentile.

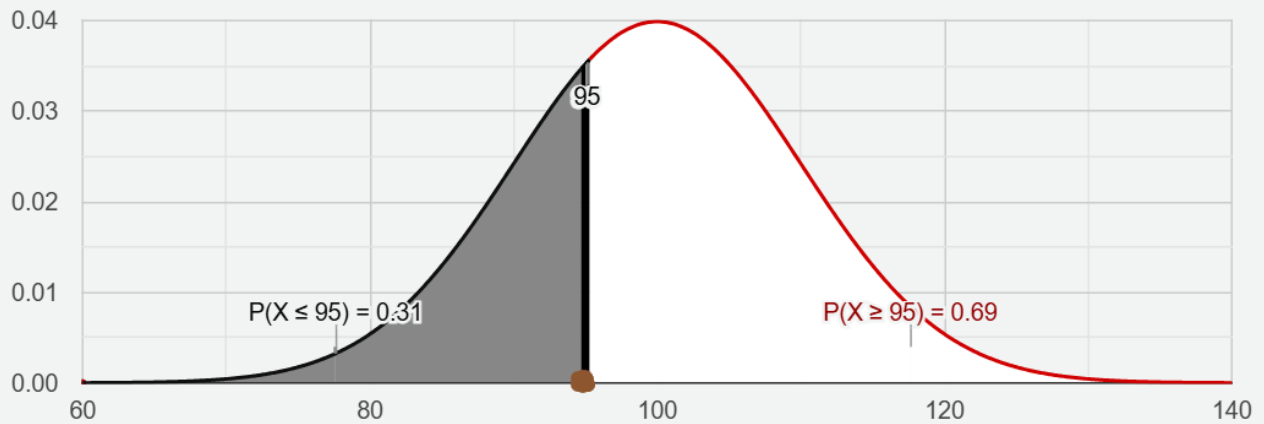


Figure  $(\text{PageIndex}\{1\})$ : Normal Curve with Area

As we learned from the previous section, a *standard normal random variable* is a normally distributed random variable with mean  $(\mu = 0)$  and standard deviation  $(\sigma = 1)$ .

It will always be denoted by the letter  $(Z)$ . The density curve for a standard normal random variable is shown in Figure  $(\text{PageIndex}\{2\})$ , and it is called the standard normal curve.

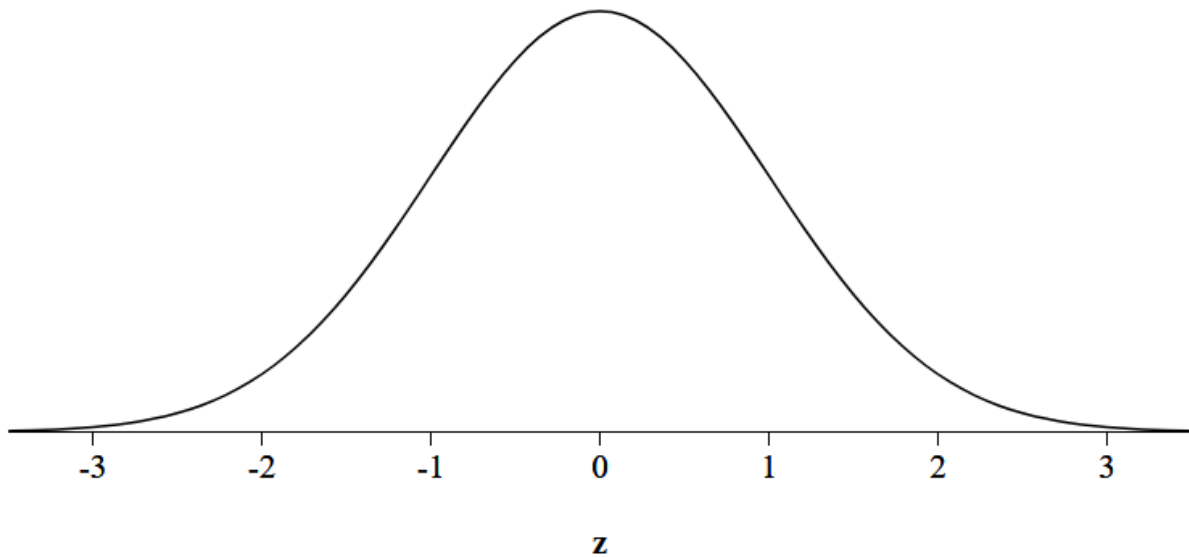


Figure  $\backslash\backslash\text{PageIndex}\{2\}\backslash$ : Standard normal curve with mean  $\backslash\backslash\mu=0\backslash$  and  $\backslash\backslash\sigma=1\backslash$

To compute probabilities for  $\backslash(z\backslash)$ , we use the cumulative probability table below. We can also use an **Online Statistics Calculator** to find the probabilities.

#### Key Point: Do Always!

When you are finding probability with normal (**or standard**) models, the first thing is to sketch the normal (**or standard normal**) curve with a label, and shade the region of interest. The region here is the area under the curve. Area could be the right area from  $\backslash(x\backslash)$  or  $\backslash(z\backslash)$  value, the left area from  $\backslash(x\backslash)$  or  $\backslash(z\backslash)$  value, or the middle area between two  $\backslash(x\backslash)$  values or  $\backslash(z\backslash)$  values.

## Finding Probability of Standard Normal Random Variable

### Finding Probability: Use Either Table or Technology

#### By Hand Using a Table

1. Draw a standard normal curve and shade the desired area.
2. Find the area to the left using the cumulative probability distribution table.

#### By Technology

It is shown in the video at the bottom of the page or explained in the example.

### Online Statistic Calculator

[Online Statistics Calculator](#)

[GeoGebra Calculator](#)

### Example $\text{\PageIndex{1}}$ : Finding Probability

Find the probabilities indicated, where as always  $Z$  denotes a standard normal random variable.

$P(Z < 1.48)$  and  $P(Z > -0.25)$ .

Answer available at the end of this section.

### Your Turn $\text{\PageIndex{1}}$ : Drag and Find the Probability

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Finding Probability: Normal Random Variable

### Finding Probability: Use either Table or Technology

If  $X$  is any normally distributed random variable, we can find the probability using the following steps.

By hand

1. Draw a normal curve and shade the desired area.
2. Find the z-score of the data value  $x$ :
3. Find the area to the left using the cumulative probability distribution table.

By Technology

It is shown in the video at the bottom of this page. Or shown in each of the following examples.

## Standard Normal Table \(\backslash\PageIndex{1}\backslash\)




Cumulative Probability  $P(Z \leq z)$

| z    | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| -3.9 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| -3.8 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| -3.7 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| -3.6 | 0.0002 | 0.0002 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| -3.5 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 |
| -3.4 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0002 |
| -3.3 | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0003 |
| -3.2 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0005 |
| -3.1 | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0007 | 0.0007 |
| -3.0 | 0.0011 | 0.0011 | 0.0011 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| -2.9 | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| -2.8 | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| -2.7 | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| -2.6 | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| -2.5 | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| -2.4 | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| -2.3 | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| -2.2 | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| -2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| -2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| -1.9 | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| -1.8 | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| -1.7 | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |
| -1.6 | 0.0548 | 0.0537 | 0.0526 | 0.0516 | 0.0505 | 0.0495 | 0.0485 | 0.0475 | 0.0465 | 0.0455 |
| -1.5 | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0721 | 0.0708 | 0.0694 | 0.0681 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| -1.1 | 0.1357 | 0.1335 | 0.1314 | 0.1292 | 0.1271 | 0.1251 | 0.1230 | 0.1210 | 0.1190 | 0.1170 |
| -1.0 | 0.1587 | 0.1562 | 0.1539 | 0.1515 | 0.1492 | 0.1469 | 0.1446 | 0.1423 | 0.1401 | 0.1379 |
| -0.9 | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |
| -0.8 | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| -0.7 | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2206 | 0.2177 | 0.2148 |
| -0.6 | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| -0.5 | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |
| -0.4 | 0.3446 | 0.3409 | 0.3372 | 0.3336 | 0.3300 | 0.3264 | 0.3228 | 0.3192 | 0.3156 | 0.3121 |
| -0.3 | 0.3821 | 0.3783 | 0.3745 | 0.3707 | 0.3669 | 0.3632 | 0.3594 | 0.3557 | 0.3520 | 0.3483 |
| -0.2 | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 | 0.3936 | 0.3897 | 0.3859 |
| -0.1 | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 | 0.4325 | 0.4286 | 0.4247 |
| 0.0  | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |

Cumulative Probability  $P(Z \leq z)$

| z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| 3.0 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |
| 3.1 | 0.9990 | 0.9991 | 0.9991 | 0.9991 | 0.9992 | 0.9992 | 0.9992 | 0.9992 | 0.9993 | 0.9993 |
| 3.2 | 0.9993 | 0.9993 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9995 | 0.9995 | 0.9995 |
| 3.3 | 0.9995 | 0.9995 | 0.9995 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9997 |
| 3.4 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9998 |
| 3.5 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 |
| 3.6 | 0.9998 | 0.9998 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.7 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.8 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

## Example \(\backslash\PageIndex{2}\backslash\): Calculate Probability Using Normal Distribution

Let  $(X)$  be a normal random variable with mean  $(\mu = 10)$  and standard deviation  $(\sigma = 2.5)$ . Compute the following probabilities.

- $P(X < 14)$
- $P(8$

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

## Your Turn \(\backslash\PageIndex{2}\backslash\): Calculate Probability Using Normal Distribution

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Example \(\backslash\PageIndex{3}\backslash\): Calculate Probability Using Normal Distribution

The lifetimes of the tread of a certain automobile tire are normally distributed with mean  $(37,500)$  miles and standard deviation  $(4,500)$  miles. Find the probability that the tread life of a randomly selected tire will be between  $(30,000)$  and  $(40,000)$  miles.

Answer available at the end of this section.

### Your Turn $\backslash(\backslash\text{PageIndex}\{3\}\backslash)$ : Find the Probability Using Normal Distribution

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Find the x-value (Data) or z-score from the Given Area

Sometimes we need to find the value of a normal random variable (or standard normal) corresponding to a given proportion, probability, area, or percentile. Look at the following graph. What is the value of  $\backslash(x)\backslash$  or  $\backslash(z)\backslash$  in the given graph?

### Percentile: How to find it Using Technology?

If  $\backslash(k\%)\backslash$  of the area under the normal curve is less than  $\backslash(P_{\{k\}})\backslash$ , then  $\backslash(P_{\{k\}})\backslash$  is called  $\backslash(k^{\text{th}})\backslash$  percentile.

Figure  $\backslash(\backslash\text{PageIndex}\{1\}\backslash)$ ,  $\backslash(115)\backslash$  is called  $\backslash(93^{\text{rd}})\backslash$  percentile.

Because area is on the left of  $\backslash(115)\backslash$ . Or  $\backslash(93\%)\backslash$  data values are less than or equal to  $\backslash(115)\backslash$ .

Note that  $\backslash(P_{\{k\}})\backslash$  is also sometime denoted by  $\backslash(X_{\{k\}})\backslash$ .

In Figure  $\backslash(\backslash\text{PageIndex}\{1\}\backslash)$ ,  $\backslash(P_{\{93\}}=X_{\{93\}}=115)\backslash$

You can find  $\backslash(P_{\{k\}})\backslash$  using an online statistic calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"
2. Type  $\backslash(\mu)\backslash$  and  $\backslash(\sigma)\backslash$  that are given in the question.
3. Check two boxes  $\backslash(P(x$
4. Type area  $\backslash(k\%)\backslash$  (in the decimal format), the number in the bottom right box. The number you will see in the left box of area is the desired percentile.

### Your Turn $\backslash(\backslash\text{PageIndex}\{4\}\backslash)$ : Find the Percentile

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Three Different Types of Shaded Areas

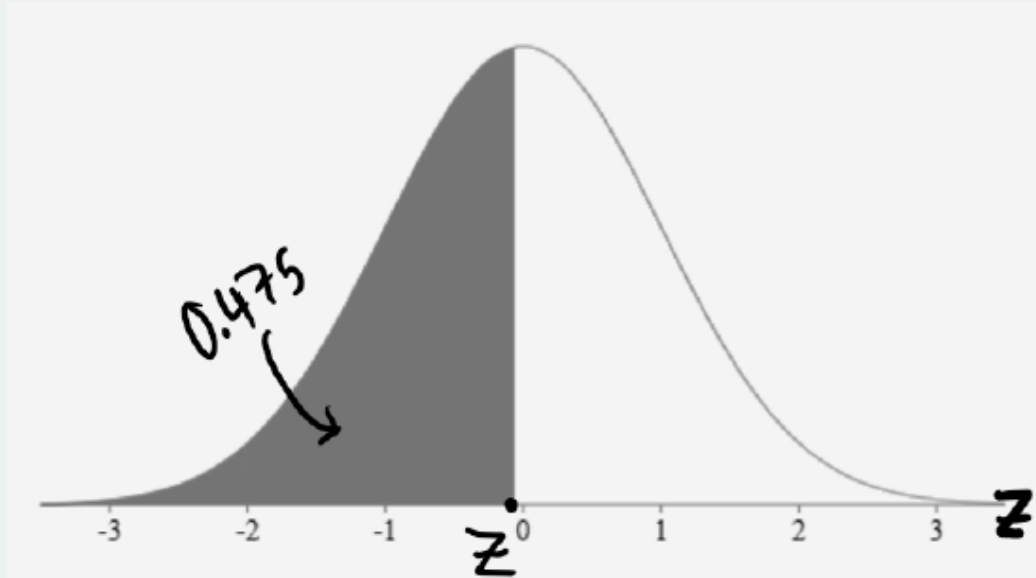


Figure  $\backslash(\backslash\text{PageIndex}\{8\}\backslash)$ :  $\backslash(P(z$

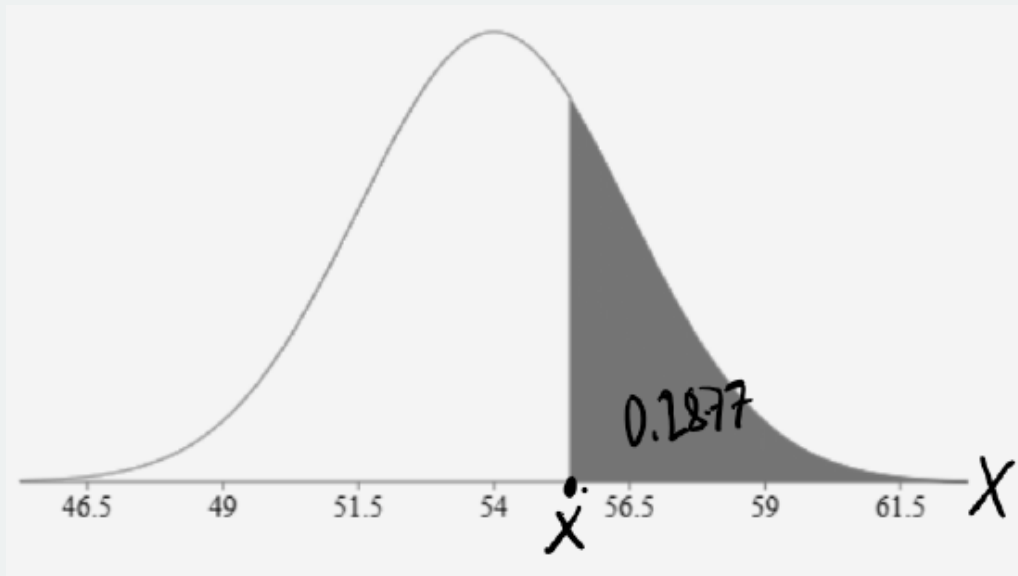


Figure  $\backslash(\backslash\text{PageIndex}\{9\}\backslash)$ :  $\backslash((P(x>a)=0.2877\backslash)$

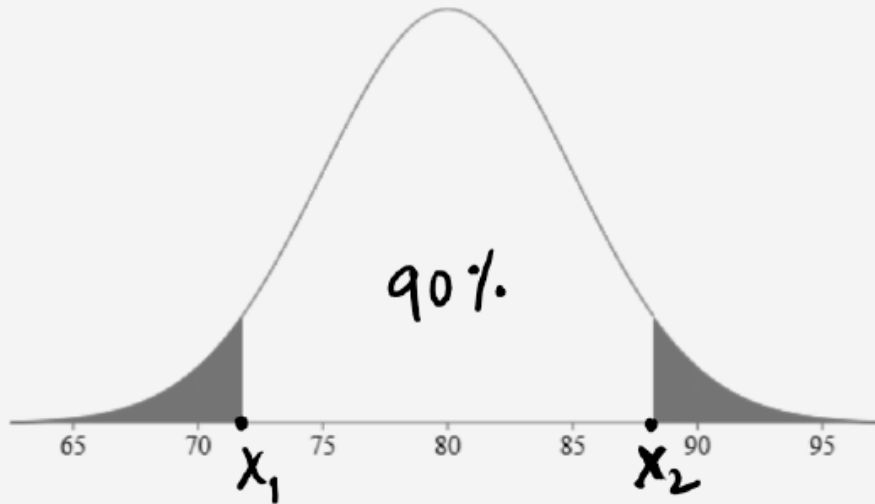


Figure  $\backslash\backslash\text{PageIndex}{10}\backslash\backslash$ :  $\backslash\backslash(P(a$

### Example $\backslash\backslash\text{PageIndex}{5}\backslash\backslash$ : Find Normal Random Variable or Standard Normal Variable

1. Figure  $\backslash\backslash\text{PageIndex}{8}\backslash\backslash$ , what is the  $\backslash\backslash(z)\backslash\backslash$ -score where the area on the left is the form that  $\backslash\backslash(z)\backslash\backslash$  is  $\backslash\backslash(0.475)\backslash\backslash$ ?
2. Figure  $\backslash\backslash\text{PageIndex}{9}\backslash\backslash$ , what is the  $\backslash\backslash(x)\backslash\backslash$  value where the area on the right is the form that  $\backslash\backslash(x)\backslash\backslash$  is  $\backslash\backslash(0.2877)\backslash\backslash$ ?
3. Figure  $\backslash\backslash\text{PageIndex}{10}\backslash\backslash$ , what are the  $\backslash\backslash(x)\backslash\backslash$  values that separate the middle  $\backslash\backslash(90\%\backslash\backslash)$  of the areas from the areas in the tails of the normal curve?

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

Answer 3 available at the end of this section.

### Your Turn $\backslash\backslash\text{PageIndex}{5}\backslash\backslash$ : Find $\backslash\backslash(z)\backslash\backslash$ socre

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Alpha $\backslash\backslash\alpha\backslash\backslash$

In a normal (or standard normal) distribution, alpha ( $\backslash\backslash\alpha\backslash\backslash$ ), is the area of just the right tail. So the value that separate area  $\backslash\backslash\alpha\backslash\backslash$  to the right to the area  $\backslash\backslash(1-\alpha)\backslash\backslash$  is called  $\backslash\backslash(z_{\alpha})\backslash\backslash$ . See the figure  $\backslash\backslash\text{PageIndex}{14}\backslash\backslash$  below.

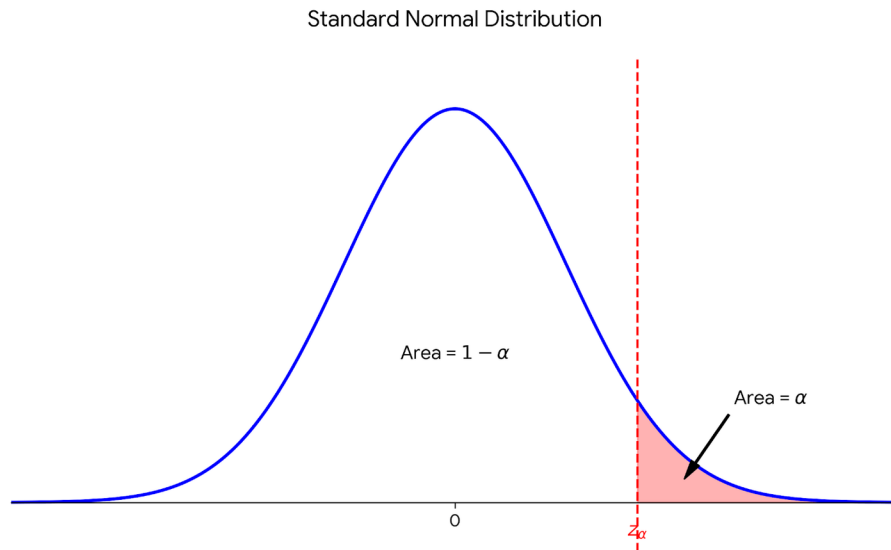


Figure  $\backslash(\backslash\text{PageIndex}\{14\}\backslash)$ : Right area  $\backslash(\backslash\alpha\backslash)$

### Your Turn $\backslash(\backslash\text{PageIndex}\{6\}\backslash)$ : Find $\backslash(z_{\backslash\alpha}\backslash)$

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example $\backslash(\backslash\text{PageIndex}\{7\}\backslash)$ : Calculate Normal Probability and Percentile

Scores on a standardized college entrance examination (*CEE*) are normally distributed with a mean of  $\backslash(510\backslash)$  and a standard deviation of  $\backslash(60\backslash)$ . A selective university considers for admission only applicants with *CEE* scores over  $\backslash(650\backslash)$ .

1. Find the percentage of all individuals who took the *CEE* who meet the university's *CEE* requirement for consideration for admission.
2. Find  $\backslash(80^{\backslash\text{th}}\backslash)$  percentile. That is the score that separates the bottom  $\backslash(80\%\backslash)$  from the top  $\backslash(20\%\backslash)$ .

*Answer 1 available at the end of this section.*

### Your Turn $\backslash(\backslash\text{PageIndex}\{8\}\backslash)$ : Calculate Normal Probability and Percentile

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Videos

### Find the Area Under the Standard Normal Curve

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



### Find X-value or Percentile Normal Curve

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



### Find the z-score when the middle area is given

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this YouTube Video online](#)



---

## Answers

### Example $\text{\PageIndex{1}}$ : Finding Probability

For  $P(z < 1.48)$ ,

Figure  $\text{\PageIndex{3}}$  shows how this probability is read directly from the table without any computation required. The digits in the ones and tenths places of  $1.48$ , namely  $1.4$ , are used to select the appropriate row of the table; the hundredths part of  $1.48$ , namely  $0.08$ , is used to select the appropriate column of the table. The four decimal place number in the interior of the table that lies in the intersection of the row and column selected,  $0.9306$ , is the probability sought:

$$P(z < 1.48) = 0.9306 \text{ \nonumber}$$

### Using an Online Statistics Calculator

You can find  $P(z < 1.48)$  using an online statistics calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu=0$  and  $\sigma=1$
3. Check two boxes  $P(x$
4. Type  $P(-\text{\textit{infinity}}$

$$P(z < 1.48) = 0.9306 \text{ \nonumber}$$

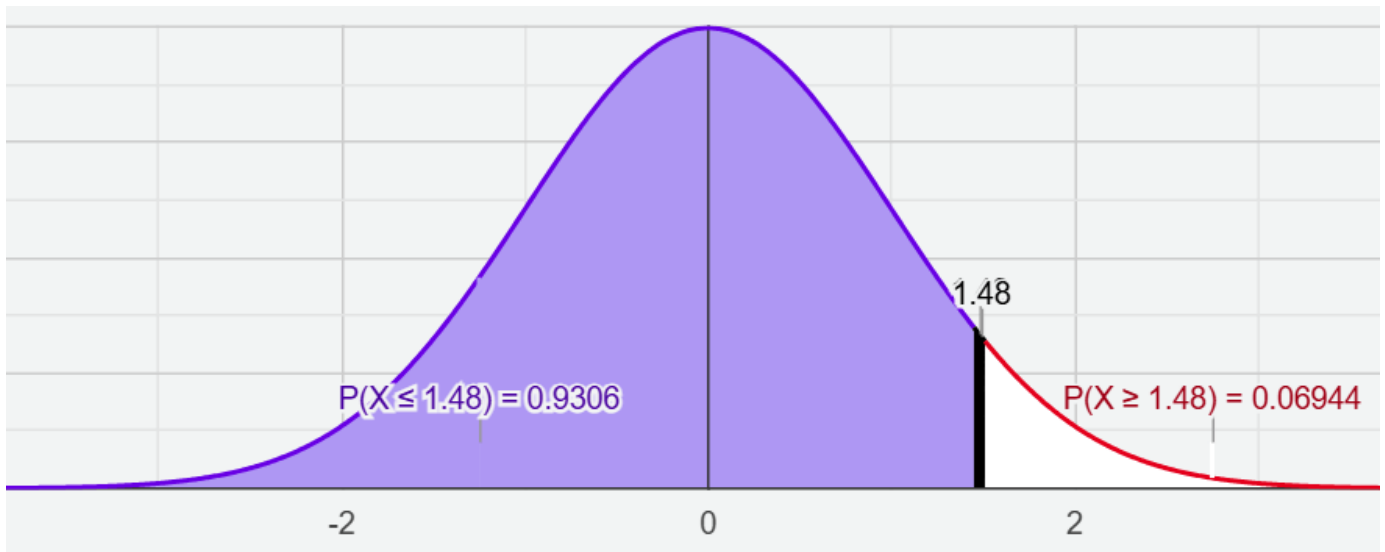


Figure 3: Computing Probabilities Using the Cumulative Table

For  $P(z > -0.25)$ ,

The minus sign in  $(-0.25)$  makes no difference in the procedure; the table is used in exactly the same way as in part (a): the probability sought is the number that is in the intersection of the row with heading  $(-0.2)$  and the column with heading  $(0.05)$ , the number  $(0.4013)$ . Thus  $P(z < -0.25) = 0.4013$ . So

$$P(z > -0.25) = 1 - 0.4013 = 0.5987$$

#### Using an Online Statistics Calculator

You can find  $P(z > -0.25)$  using an online statistic calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"
2. Type  $(\mu=0)$  and  $(\sigma=1)$
3. Check two boxes  $(P(x > a))$  and Calculate Area
4. Type  $P(-0.25)$

$$P(z > -0.25) = 0.5987$$

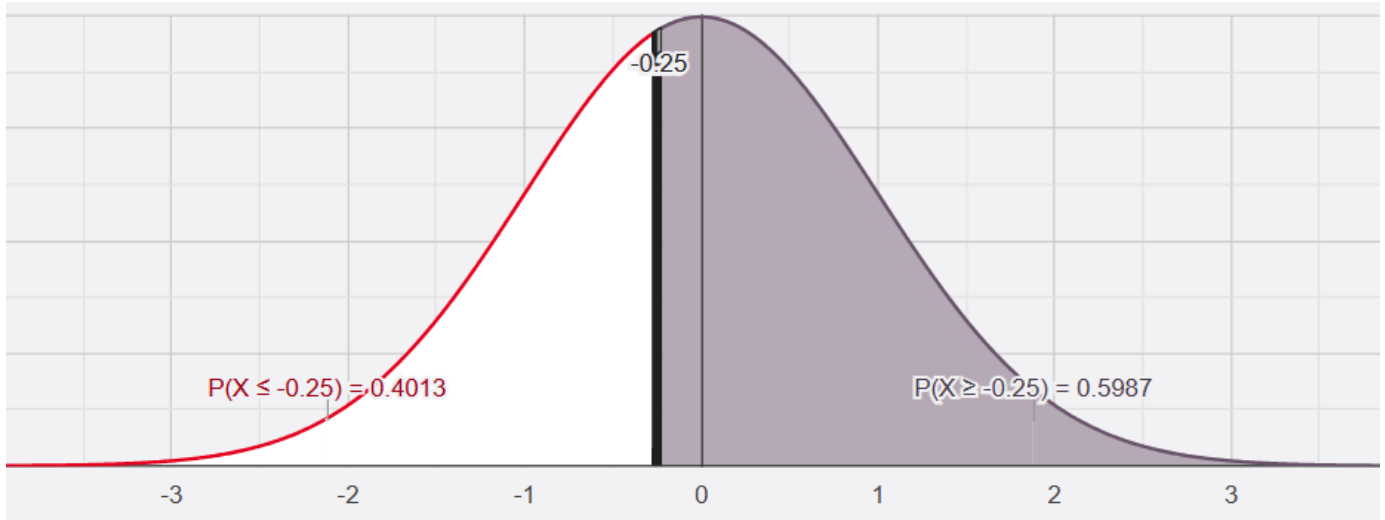


Figure 4: Computing Probabilities Using the Cumulative Table or Online Calculator

#### Example 2: Calculate Probability Using Normal Distribution

$z$  score of 14 is

$$z = \frac{14 - 10}{2.5} = 1.6$$

So

$$P(x < 14) = P(z < 1.6) = 0.9452$$

Now look at the table to find the probability (area).

And the normal curve with the shaded area represents the desired area in figure 5.

Or you can find  $P(x < 14)$  using an online statistic calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"

2. Type  $\mu=10$  and  $\sigma=2.5$
3. Check two boxes  $P(x$
4. Type  $P(-\infty)$

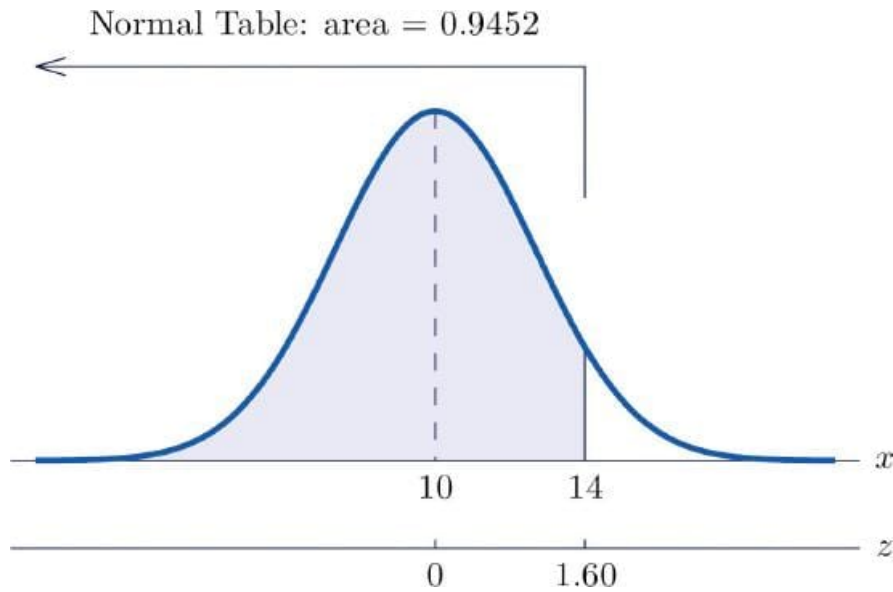


Figure  $\backslash(\backslash\text{PageIndex}{5})\backslash$ : Probability Computing for a Normal Random Variable

**Example  $\backslash(\backslash\text{PageIndex}{2})\backslash$ : Calculate Probability Using Normal Distribution**

See Figure  $\backslash(\backslash\text{PageIndex}{4})\backslash$  "Probability Computation for a General Normal Random Variable".  $z$  score of  $(8)$  is

$$z = \frac{8 - 10}{2.5} = -0.80$$

$z$  score of  $(14)$  is

$$z = \frac{14 - 10}{2.5} = 1.6$$

And the probability between two values is given by

$$P(8$$

And the normal curve with the shaded area represents the desired area in figure  $\backslash(\backslash\text{PageIndex}{6})\backslash$ .

Or you can find  $P(8$

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu=10$  and  $\sigma=2.5$
3. Check two boxes  $P(a$
4. Type  $P(8$

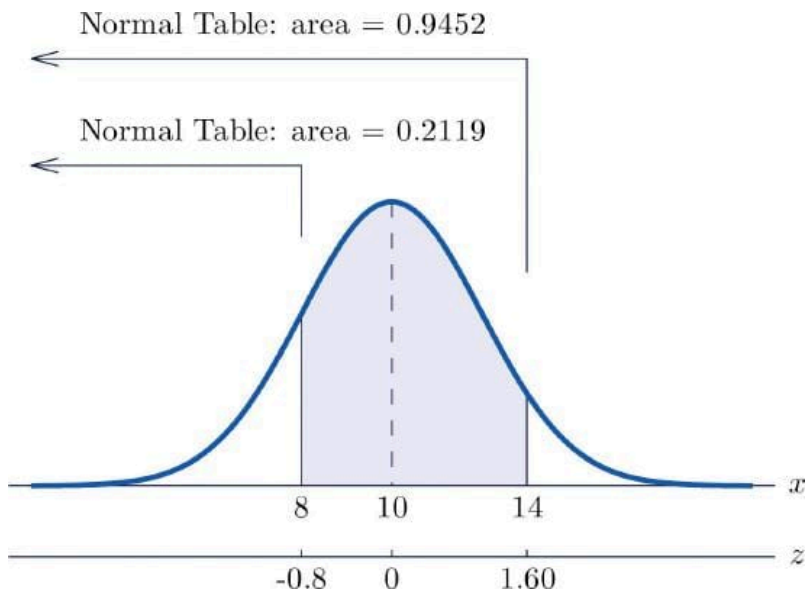


Figure  $\backslash(\backslash\text{PageIndex}{6})\backslash$ : Probability Computing for a Normal Random Variable

**Example  $\backslash(\backslash\text{PageIndex}{3})\backslash$ : Calculate Probability Using Normal Distribution**

Let  $(X)$  denote the tread life of a randomly selected tire. To make the numbers easier to work with, we will choose thousands of miles as the units. Thus  $\mu = 37,500$ ,  $\sigma = 4,500$ , and the problem is to compute  $P(30,000$

$z$ -score of 30,000 is

$$z = \frac{30,000 - 37,500}{4,500} = -1.67$$

$z$ -score of 40,000 is

$$z = \frac{40,000 - 37,500}{4,500} = 0.56$$

And the probability between two values is given by

$$P(30,000$$

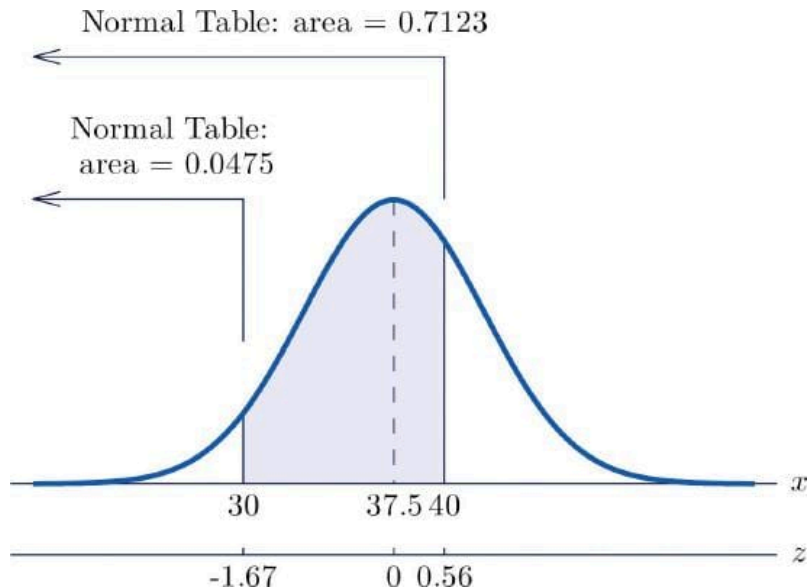


Figure 7: Probability Computation for Tire Tread Wear

Note that the two  $z$ -scores were rounded to two decimal places in order to use the "Cumulative Normal Probability" table.

Or you can use the technology directly: see the attached video at the bottom of the page.

#### Example 5: Find Normal Random Variable or Standard Normal Variable

We want to find the  $z$ -score such that the area to the left of that  $z$ -score is 0.475.

Since the curve is a standard normal curve, put  $\mu=0$  and  $\sigma=54-51.5=2.5$ . We use an Online Statistic Calculator to find it.

So the  $z$ -score is -0.0627. See the attached figure 11.

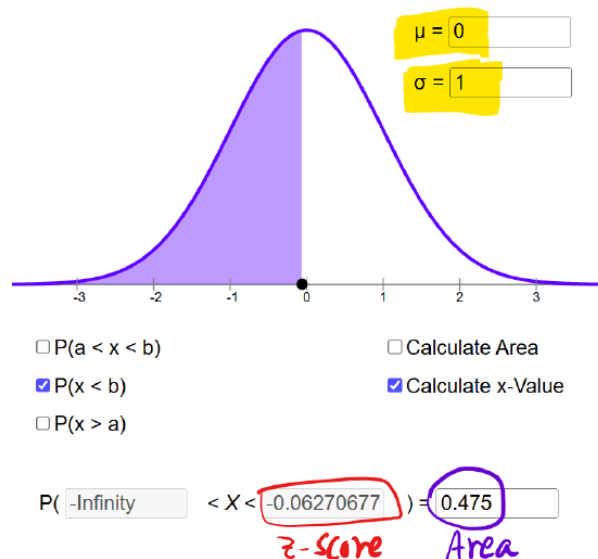


Figure 11:  $z$ -score from the area on the left of that  $z$ -score is 0.475

#### Example 5: Find Normal Random Variable or Standard Normal Variable

We want to find the  $x$  value such that the area on the right of that value is 0.2877.

Since the curve is a normal curve, put  $\mu=54$  and  $\sigma=54-51.5=2.5$ .

We use an Online Statistics Calculator to find it. So  $x$  value is  $55.4$ . See the attached  $\text{\PageIndex{12}}$ .

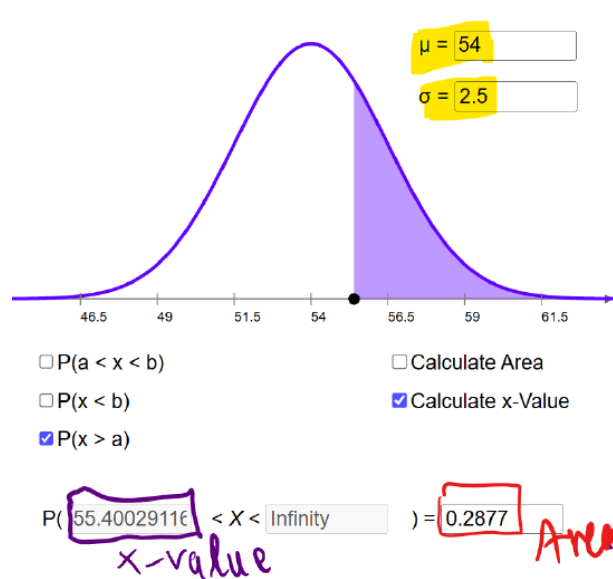


Figure  $\text{\PageIndex{12}}$ :  $x$  value from the area on the right is  $0.2877$

#### Example $\text{\PageIndex{5}}$ : Find Normal Random Variable or Standard Normal Variable

We want to find the  $x$  values such that the area between the two values is  $0.90$ . Since the curve is a normal curve, put  $\mu=80$  and  $\sigma=80-75=5$ . We use an Online Statistics Calculator to find it. So  $x$  values are  $71.78$  and  $88.22$ . See the attached  $\text{\PageIndex{13}}$ . Note that  $71.78$  is  $5$ th percentile and  $88.22$  is  $95$ th percentile.

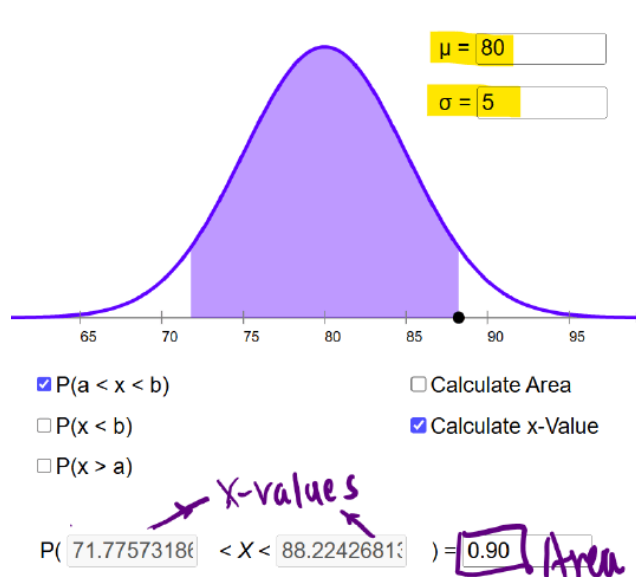


Figure  $\text{\PageIndex{13}}$ :  $x$ -values, the area on the left of those two values is  $0.90$

#### Example $\text{\PageIndex{7}}$ : Calculate Normal Probability and Percentile

##### Using Table:

Let  $X$  denote the score made on the CEE by a randomly selected individual. Then  $X$  is normally distributed with mean  $510$  and standard deviation  $60$ . The probability that  $X$  lies in a particular interval is the same as the proportion of all exam scores that lie in that interval.

Thus, the solution to the problem is  $P(X > 650)$ , expressed as a percentage. Figure  $\text{\PageIndex{12}}$  "Probability Computation for Exam Scores" illustrates the following computation:  $z = \frac{650-510}{60} = 2.33$

$$P(x > 650) = P(Z > 2.33) = 1 - 0.9901 = 0.0099$$

The proportion of all CEE scores that exceed  $650$  is  $0.0099$ , hence  $0.99\%$  or about  $1\%$ .

##### Using an Online Statistics Calculator

You can find  $P(x > 650)$  using an online statistic calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu=650$  and  $\sigma=60$
3. Check two boxes  $P(x > a)$  and Calculate Area
4. Type  $P(650)$

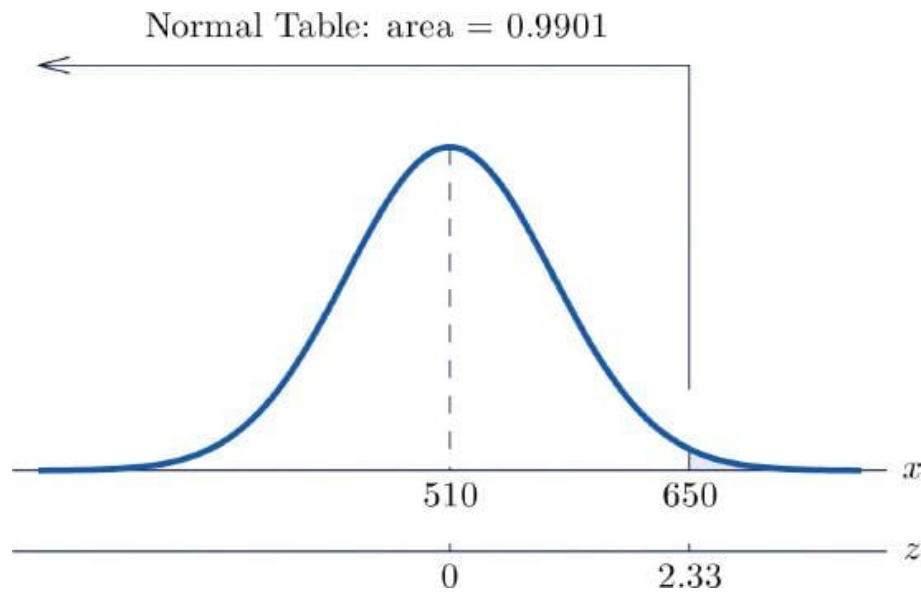


Figure \(\PageIndex{15}\): Probability Computation for Exam Scores

\(80\%\\) percentile, \(\mathcal{P}\_{80}\)\) is score that separate top \(\mathcal{20\%}\\) score to bottom \(\mathcal{80\%}\\) score.

#### Using an Online Statistics Calculator

You can find \(\mathcal{P}\_{80}\)\) using an online statistic calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"
2. Type \(\mu=510\)\) and \(\sigma=60\)\)
3. Check two boxes \(\mathcal{P}(x\)
4. Type area \(\mathcal{0.80}\)\), the number in the bottom right box. The number you will see in the left box of \(\mathcal{0.80}\)\) is the desired percentile.

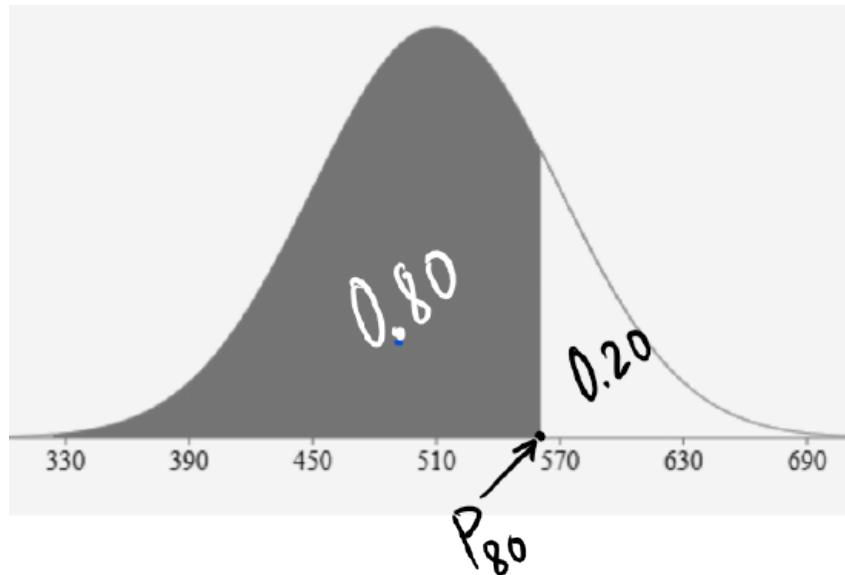


Figure \(\PageIndex{16}\): Showing \(\mathcal{80^{th}}\)\) percentile, \(\mathcal{P}\_{80}=560.50\)\).

This page titled [6.2: Normal Distribution and Applications](#) is shared under a license and was authored, remixed, and/or curated by Anonymous via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [4.9: Hyperbolic Functions](#) by [Larry Green](#) is licensed [CC BY 4.0](#).
- [7.7: What Are the Odds?](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/contemporary-mathematics>.
- [Derivative of arcsech](#) by [Larry Green](#) is licensed [CC BY 4.0](#).

## 7: SAMPLING DISTRIBUTIONS

A statistic, such as the sample mean or the sample standard deviation, is a number computed from a sample. Since a sample is random, every statistic is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. As a random variable it has a mean, a standard deviation, and a probability distribution. The probability distribution of a statistic is called its sampling distribution. Typically sample statistics are not ends in themselves, but are computed in order to estimate the corresponding population parameters. This chapter introduces the concepts of the mean, the standard deviation, and the sampling distribution of a sample statistic, with an emphasis on the sample mean

### 7.1: The Sampling Distribution of the Sample Mean

### 7.2: The Sampling Distribution of Sample Proportion

---

This page titled [7: Sampling Distributions](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by LibreTexts.

- [1.5: Surface Area of Revolution](#) by [Larry Green](#) is licensed [CC BY 4.0](#).

## 7.1: The Sampling Distribution of the Sample Mean

### Introduction

In inferential statistics, we use sample characteristics (statistic) to estimate population characteristics (parameter).

We use sampling distributions because, in real-world situations, we usually do not have enough time, money, or resources to measure an entire population. Instead, we rely on samples to make conclusions about the population.

Even though individual data values may vary, the sampling distribution helps us understand how sample means behave and how close they are likely to be to the true population mean.

For example, suppose you want to estimate the average age of all students at **Florida SouthWestern State College**. Since it would be difficult to ask every student, you instead take a random sample of students and calculate the sample mean.

To understand the behavior of sample means, imagine the following process:

1. Randomly select a sample of 50 students from the campus.
2. Calculate the average age of that sample.
3. Repeat this process many times, recording the sample mean each time.

If you collect all of those sample means, this is called the **sampling distribution of the sample mean (SDOSM)**. You also graph (make a histogram) of all the means.

#### Notation Regarding SDOSM

Population mean =  $\mu$

Sample mean =  $\bar{x}$

Population standard deviation =  $\sigma$

Sample standard deviation =  $s$

Sample size =  $n$

Population size =  $N$

Mean of the sampling distribution of the sample mean =  $\mu_{\bar{x}}$

Standard deviation of the sampling distribution of the sample mean =  $\sigma_{\bar{x}}$

Consider the population data with size ( $N = 5$ ). Data is the commute time (in minutes) of 5 students to school

$X : 10, 20, 5, 30, 15$

Let's choose a sample of size  $n = 3$ .

All possible sample sizes of 3 from the population

Sample 1 : 10, 20, 5 sample mean  $\bar{x} = 11.67$

Sample 2 : 10, 20, 30 sample mean  $\bar{x} = 20$

Sample 3 : 10, 20, 15 sample mean  $\bar{x} = 15$

Sample 4 : 10, 30, 5 sample mean  $\bar{x} = 15$

Sample 5 : 10, 5, 15 sample mean  $\bar{x} = 10$

Sample 6 : 10, 30, 15 sample mean  $\bar{x} = 18.33$

Sample 7 : 5, 20, 30 sample mean  $\bar{x} = 18.33$

Sample 8 : 20, 15, 5 sample mean  $\bar{x} = 13.33$

Sample 9 : 30, 20, 15 sample mean  $\bar{x} = 21.67$

Sample 10 : 5, 30, 15 sample mean  $\bar{x} = 16.67$

A collection of all possible sample means is

$$\bar{X}: 1.67, 20, 15, 15, 10, 18.33, 18.33, 13.33, 21.67, 16.67$$

The mean of all sample means (Think about this as another data set)

$$\mu_{\bar{x}} = \frac{11.67 + 20 + 15 + 15 + 10 + 18.33 + 18.33 + 13.33 + 21.67 + 16.67}{10} = 16$$

The standard deviation of all sample means

$$\sigma_{\bar{x}} = 3.8$$

Actually, the population mean

$$\mu = \frac{10 + 20 + 5 + 30 + 15}{5} = 16$$

The population standard deviation is

$$\sigma = 8.6$$

### Sampling Distribution of Sample Means

Sampling distribution of the sample mean (SDOSM) is the collection of all possible sample means  $\bar{x}$  taken from the same population using samples of the same size. SDOSM would be graphed.

SDOSM is also denoted by  $\bar{X}$ .

Here  $\bar{X}$  is random variable with parameters  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$ .

### Underlying Population

The **underlying population** is the entire group of individuals, objects, or data from which samples are taken in a study. There are two possible situations regarding the distribution of the underlying POPULATION when we do a sampling distribution of the sample mean. They could have had the following distribution.

1. The population could be normal.
2. The population could not be normal or unknown.

### When Sampling Distribution of the Sample Mean (SDOSM) Approximately Normal?

The sampling distribution of the sample mean is approximately normal if:

- The population is normal, or
- the sample size is large  $n \geq 30$

So the shape of the underlying population affects the sampling distribution:

- If the underlying population is normal, the sampling distribution of sample mean is normal.
- If the sample size is large  $\geq 30$ , the sampling distribution is approximately normal even if the underlying population is not normal.

The mean and standard deviation of the population in the above example are  $\mu = 16$  and  $\sigma = 8.6$ . The mean of the sample mean  $\bar{x}$  that we have just computed is exactly the mean of the population. The standard deviation of the sample mean  $\bar{x}$  that we have just computed is the standard deviation of the population divided by the square root of the sample size:  $\sigma_{\bar{x}} = \frac{8.6}{\sqrt{5}}$ . These relationships

are not coincidences, but are illustrations of the following formulas.

## Central Limit Theorem (CLT)

This concept is so important and plays such a critical role in what follows it deserves to be developed further. Indeed, there are two critical issues that flow from the Central Limit Theorem and the application of the Law of Large Numbers to it. These are

1. The probability density function of the sampling distribution of means is normally distributed **regardless** of the underlying distribution of the population observations The
2. The standard deviation of the sampling distribution decreases as the sample size used to calculate the sampling distribution means increases.

Taking these in order. It would seem counterintuitive that the population may have **any** distribution, and the distribution of means coming from it would be normally distributed. Using computers, experiments can be simulated to show how the sampling distribution changes as the sample size increases. These simulations show visually the results of the mathematical proof of the Central Limit Theorem.

### CLT for Normal Population

The following graphs show that the SDOSM is normally distributed when the population is normally distributed.

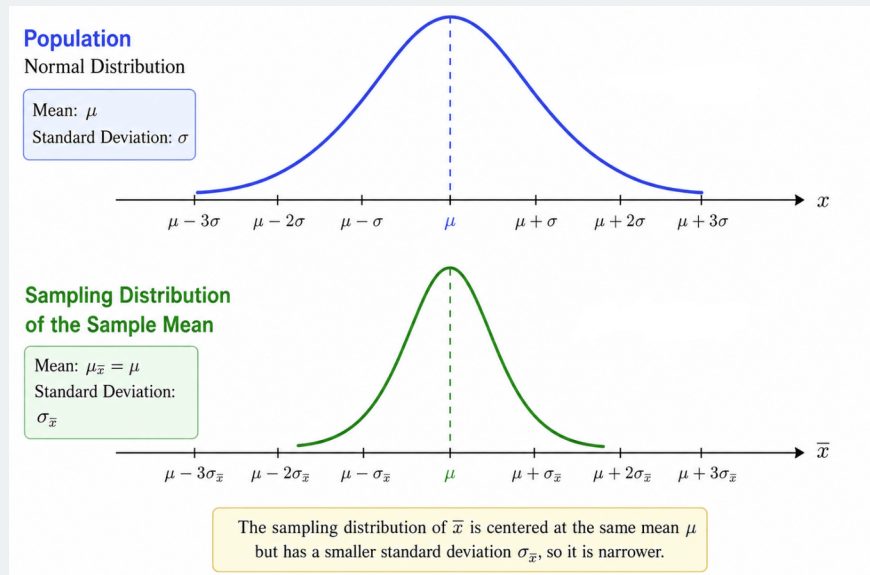


Figure 7.1.1: Distribution of  $x$  and  $\bar{x}$

Here are three examples of very different population distributions, along with the evolution of their sampling distributions toward normality as the sample size increases. The top panel in these cases represents the histogram for the original data. The three panels show the histograms for 1,000 randomly drawn samples for different sample sizes:  $n = 10$ ,  $n = 25$  and  $n = 50$ . As the sample size increases while the number of samples remains constant, the distribution of the 1,000 sample means approaches the smooth line representing the normal distribution.

Figure 7.1.2 is for a normal distribution of individual observations, and we would expect the sampling distribution to converge on the normal quickly. The results show this and that, even with a very small sample size, the distribution is close to normal.

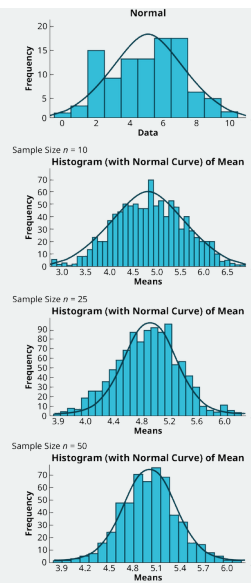


Figure 7.1.2: CLT Normal Population

### CLT for Skewed Population

Population distribution in figure 7.1.3 is skewed. For skewed distributions, our intuition would suggest that it will take larger sample sizes to approach normality, and indeed, that is what we observe in the simulation. Nevertheless, with a sample size of 50, not considered very large, the distribution of sample means has clearly taken on the shape of the normal distribution.

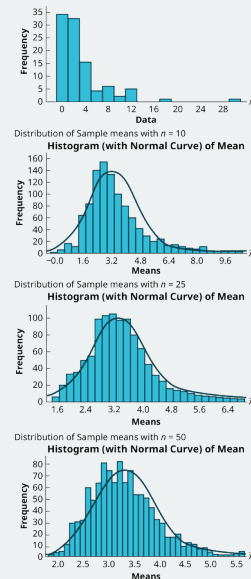


Figure 7.1.3: CLT for Skewed Distribution

The Central Limit Theorem provides more than the proof that the sampling distribution of means is normally distributed. It also provides us with the mean and standard deviation of this distribution. Further, as discussed above, the expected value of the mean,  $\mu_{\bar{x}}$ , is equal to the mean of the population of the original data, which is what we are interested in estimating from the sample we took. And standard deviation is  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ , and this is critical to calculate probabilities of values of the new random variable,  $\bar{x}$ .

Figure 7.1.4 shows three sampling distributions. The only change that was made is the sample size that was used to get the sample means for each distribution. As the sample size increases,  $n$  goes from 10 to 30 to 50, the standard deviations of the respective sampling distributions decrease because the sample size is in the denominator of the standard deviations of the sampling distributions.

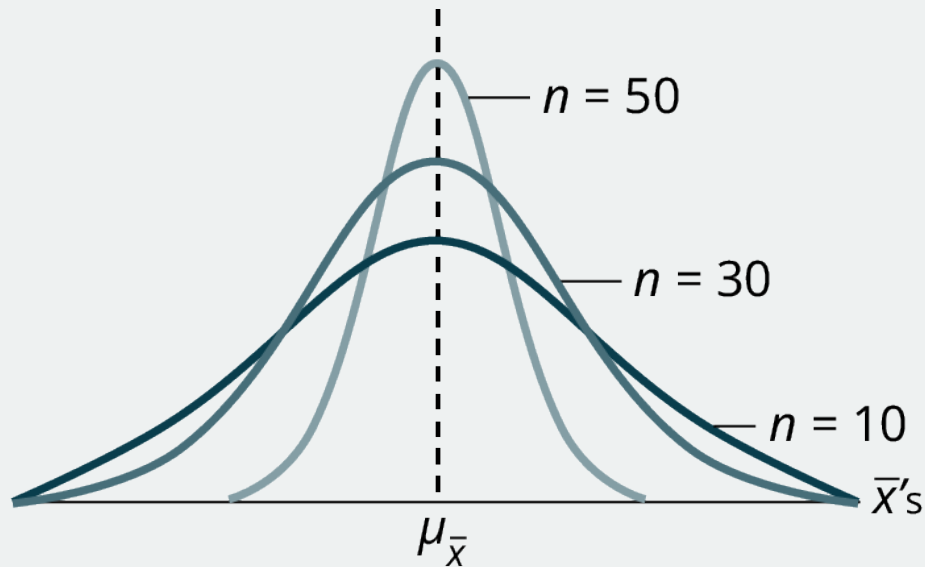


Figure 7.1.4: Showing Three Sampling Distributions

#### Definition: Mean of SDOSM and Standard Deviation of SDOSM

Suppose random samples of size  $n$  are drawn from a population with mean  $\mu$  and standard deviation  $\sigma$ . The mean  $\mu_{\bar{x}}$  and standard deviation  $\sigma_{\bar{x}}$  of the sample mean of sampling distribution of the sample mean are given by

$$\mu_{\bar{x}} = \mu$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

As we see, if we could take every possible sample from the population and compute the corresponding sample mean, those numbers would center on the population mean  $\mu$ .

$\sigma_{\bar{x}}$  is called the standard error of the mean.

Note:

1.  $x$ , the measurement of a single element selected at random from the population; the distribution of  $x$  is the distribution of the population, with mean the population mean  $\mu$  and standard deviation the population standard deviation  $\sigma$ ;
2.  $\bar{x}$ , the mean of the measurements in a sample of size  $n$ ; the distribution of  $\bar{x}$  is its sampling distribution, with mean  $\mu_{\bar{x}} = \mu$ .

#### Example 7.1.1: Population is Normal

The mean and standard deviation of the tax value of all vehicles registered in a certain state are  $\mu = \$13,525$  and  $\sigma = \$4,180$ . Suppose the value of all registered vehicles is normally distributed. Suppose random samples of size 100 are drawn from the population of vehicles.

Is the sampling distribution of the sample mean approximately normal? What are the mean  $\mu_{\bar{x}}$  and standard deviation  $\sigma_{\bar{x}}$  of the sampling distribution of sample mean?

Answer available at the end of this section.

The importance of the Central Limit Theorem (above) is that it allows us to make probability statements about the sample mean, specifically about its relationship to the population mean, as we will see in the examples. But to use the result properly, we must first realize that there are two separate random variables (random variable  $X$  and  $\bar{X}$  (and therefore two probability distributions) at play.

### Your Turn 7.1.1: Population is Normal

### Example 7.1.2: Population is not Normal

An automobile battery manufacturer claims that its mid-grade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is unknown. Suppose random samples of size 36 are drawn from the population.

Is the sampling distribution of the sample mean approximately normal? What are the mean  $\mu_{\bar{x}}$  and standard deviation  $\sigma_{\bar{x}}$  of the sampling distribution of sample mean  $\bar{x}$ ?

Answer available at the end of this section.

### Your Turn 7.1.2: Population is Skewed

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Calculating Probability Using SDOSM

When you scale up from individual data points to **sample means** (averages calculated from groups), you are working with the **sampling distribution of the sample mean**.

The math changes slightly because of a beautifully predictable rule: **groups are less variable than individuals**. If you pick one random person, they might be incredibly tall or short. But if you pick a random group of 50 people, their *average* height will almost always be pretty close to the national average.

### What do you need to look for when calculating probability using SDOSM?

When you calculate probability using the Sampling Distribution of the Sample Mean (SDOSM), make sure to check the following:

1. Is the sampling distribution of the sample mean approximately normal?
2. Which area are you finding, Or what is the area of interest? The left area, the right area, or the area between two values.
3. What is the sample size  $n$ ?
4. What are the formulas for the mean and standard deviation of the sampling distribution?

$$\mu_{\bar{x}} = \mu$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Always draw and label the normal curve first and shade the area of interest before calculating the probability.

### Parameter for $X$ and $\bar{X}$

If the population distribution is unknown or skewed, you cannot use the normal model to determine probabilities for a randomly selected individual (that is, when  $n = 1$ ). In that case, you must say the probability does not exist using the normal mode. However, if the sample size is sufficiently large  $n \geq 30$ , the sampling distribution of the sample mean may still be approximately normal by the Central Limit Theorem.

Not that

For  $X$ : Parameters are  $\mu$  and  $\sigma$

For  $\bar{X}$ : Parameters are  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$

### Example 7.1.3: Calculating Probability Using Normal Model

A random sample of size 50 drawn from a population with unknown distribution with mean 112 and standard deviation 40.

1. Find the probability that a randomly selected element exceed 113, that is, compute  $P(x > 113)$ ,
2. Find the mean and standard deviation of SDOSM.
3. Find the probability that the sample mean is between 110 and 114.
4. Find the probability that the sample mean is greater than 113.

Answer 1 available at the end of this section.

### Your Turn 7.1.3: Calculating Probability Using Normal Model

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example 7.1.4: Calculating Probability Using Normal Model

A prototype automotive tire has a design life of 38,500 miles with a standard deviation of 2,500 miles. Five such tires are manufactured and tested. On the assumption that the actual population mean is 38,500 miles and the actual population standard deviation is 2,500 miles, find the probability that the sample mean will be less than 36,000 miles. Assume that the distribution of lifetimes of such tires is normal.

Answer available at the end of this section.

## Interactive Content

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Your Turn 7.1.6: Terminology for SDOSM

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Answers

### Example 7.1.1: Population is Normal

Since the population ( $X$ ) distribution is normal, regardless of the sample size ( $n = 100$ ), the sampling distribution of the sample mean is approximately normal. That is SDOSM ( $\bar{X}$ ) is normal with the following parameters.

Mean of SDOSM

$$\mu_{\bar{x}} = \mu = \$13,525$$

Standard deviation of SDOSM

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

### Example 7.1.2: Population is not Normal

Since the population ( $X$ ) distribution is unknown. But since  $n = 36 \geq 30$ , the sampling distribution of the sample mean is approximately normal. That is  $\bar{X}$  is normal with the following parameters.

Mean of SDOSM

$$\mu_{\bar{x}} = \mu = 50$$

Standard deviation of SDOSM

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1$$

### Example 7.1.3: Calculating Probability Using Normal Model

Computation of the probability  $P(x > 113)$  is not possible, since we do not know the distribution of  $X$ . We say  $P(x > 113) = DNE$ . By contrast, we could compute  $P(\bar{x} > 113)$  (in part 4) even without complete knowledge of the distribution of  $X$  because the Central Limit Theorem guarantees that  $\bar{X}$  is approximately normal.

Use the formulas to find the mean and standard deviation of SDOSM

$$\mu_{\bar{x}} = \mu = 112$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.65685$$

Since the sample size is at least 30, SDODM is approximately normally distributed.

We compute probabilities **using an Online Statistics Calculator**

You can find  $P(x > 113)$  using an online statistic calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 112$  and  $\sigma = 5.65685$
2. Check two boxes  $\backslash P(a$
3. Type  $\backslash P(113$

$$P(110 < \bar{x} < 114) = 0.2763$$

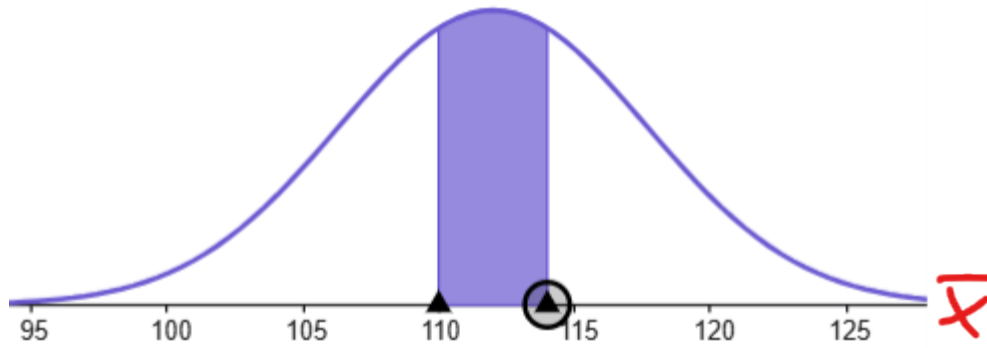


Figure 7.1.5: Shaded area is 0.2763

We compute probabilities **using an Online Statistics Calculator**

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 112$  and  $\sigma = 5.65685$
2. Check two boxes  $P(x > a)$  and Calculate Area
3. Type  $\backslash P(113$

$$P(\bar{x} > 113) = 0.4298$$

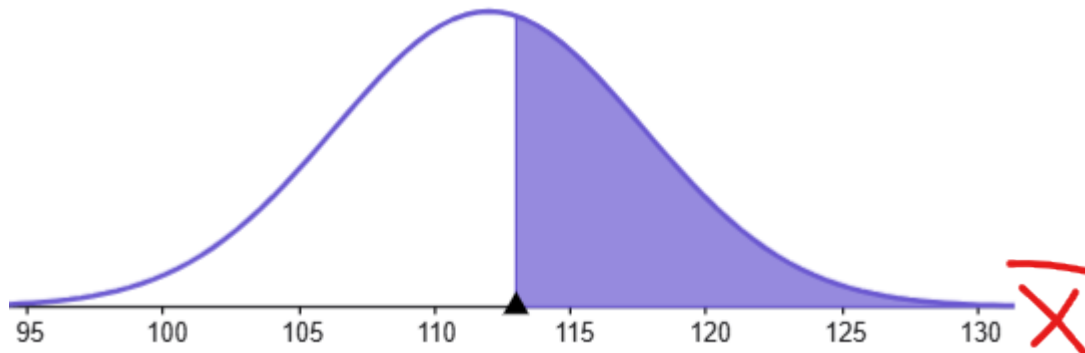


Figure 7.1.6: Shaded area is 0.4298

#### Example 7.1.4: Calculating Probability Using Normal Model

For simplicity, we use units of thousands of miles. Then the mean of  $\bar{X}$  is  $\mu_{\bar{x}} = \mu = 38,500$  and standard deviation of  $\bar{X}$  is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2,500}{\sqrt{5}} = 1118.0340$$

Since the population is normally distributed, so is  $\bar{X}$  approximately normal.

Find z-score of 36,000

$$z = \frac{36,000 - 38,500}{1118.0340} = -2.24$$

So Probability is

$$\begin{aligned} P(\bar{x} < 36,000) &= P(x < -2.24) \\ &= 0.0125 \end{aligned}$$

That is, if the tires perform as designed, there is only about a 1.25% chance that the average of a sample of this size would be so low.

We compute probabilities **using an Online Statistics Calculator**

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 38,500$  and  $\sigma = 1118.0340$
2. Check two boxes  $\backslash P(x$
3. Type  $\backslash P(-\text{infinity})$

$$P(\bar{x} < 36,000) = 0.0128$$

---

This page titled [7.1: The Sampling Distribution of the Sample Mean](#) is shared under a 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.

- [4.10: Stokes' Theorem](#) is licensed [CC BY 4.0](#).
- [Stokes' Theorem II](#) by [Larry Green](#) is licensed [CC BY 4.0](#).

## 7.2: The Sampling Distribution of Sample Proportion

### Why do we use the Sampling Distribution of the Sample Proportion?

In inferential statistics, we use sample characteristics (statistic) to estimate population characteristics (parameter). In this section,  $p$  population proportion is a parameter and sample proportion  $\hat{p}$  is a statistic.

For example, suppose you want to estimate what proportion of the TikTok app is used while on campus at **Florida SouthWestern State College**. Since it would be difficult to ask every student, you take a random sample and calculate the sample proportion.

To understand the behavior of the sample proportion, imagine the following process:

1. Randomly select a sample of 100 students from the campus.
2. Calculate the sample proportion of that sample.
3. Repeat this process many times, recording the sample proportion each time.

If you collect all of those sample proportions, this is called the **sampling distribution of the sample proportion**. You also graph (make a histogram) of all the means. That is also called the **sampling distribution of the sample proportion (SDOSP)**.

Often, sampling is done in order to estimate the proportion of a population that has a specific characteristic, such as the proportion of all items coming off an assembly line that are defective or the proportion of all people entering a retail store who make a purchase before leaving.

Craft a simple, neutral survey question to avoid response bias: "Do you use the TikTok app while on campus? (Yes / No)" Ensure the survey is anonymous so students answer honestly.

The sample proportion is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. Viewed as a random variable, it will be written  $\hat{p}$ .

#### Formula Sample Proportion

$$\hat{p} = \frac{x}{n}$$

Where  $n$  is a sample size and  $x$  is the number of individuals with given characteristics.

Note that the value of the sample proportion and the population proportion is always between 0 and 1.

Even though individual data values may vary, the sampling distribution helps us understand how sample proportions behave and how close they are likely to be to the true population proportions.

#### Notation Regarding SDOSP

Population proportion =  $p$

Sample proportion =  $\hat{p}$

Sample size =  $n$

Population size =  $N$

Mean of the sampling distribution of the sample proportion =  $\mu_{\hat{p}}$

Standard deviation of the sampling distribution of the sample proportion =  $\sigma_{\hat{p}}$

Suppose a university with 20,000 students wants to know the proportion of students who prefer coffee over tea.

- True Population Parameter:  $p = 0.60$  or 60% of all students prefer coffee.
- Sample Size:  $n$ : Each sample consists of  $n = 100$  randomly selected students.

Five different student researchers independently conduct surveys of 100 students across campus.

Table 7.2.1: Sampling Distribution of Sample Proportion

|                 | Location sampled | Sample Size ( $n$ ) | Students preferring coffee ( $x$ ) | Sample Proportion                 |
|-----------------|------------------|---------------------|------------------------------------|-----------------------------------|
| <b>Sample 1</b> | Student Union    | 100                 | 63                                 | $\hat{p} = \frac{63}{100} = 0.63$ |
| <b>Sample 2</b> | Library          | 100                 | 58                                 | $\hat{p} = \frac{58}{100} = 0.58$ |
| <b>Sample 3</b> | Science Building | 100                 | 65                                 | $\hat{p} = \frac{65}{100} = 0.65$ |
| <b>Sample 4</b> | Dining Hall      | 100                 | 55                                 | $\hat{p} = \frac{55}{100} = 0.55$ |
| <b>Sample 5</b> | Rec Center       | 100                 | 61                                 | $\hat{p} = \frac{61}{100} = 0.61$ |

A collection of all possible sample means that we collected is

$$\hat{p} : 0.63, 0.58, 0.65, 0.55, 0.61$$

Here  $\hat{p}$  is called a random variable. Because it varies from sample to sample.

This variation of the sample proportion from sample to sample is called **sampling variability**.

The mean of all sample proportions (Think about this as another data set)

$$\mu_{\hat{p}} = \frac{0.63 + 0.58 + 0.65 + 0.55 + 0.61}{5} = 0.604 \approx 0.60$$

Notice that no single sample got exactly 0.60 (60%), but all samples clustered around the true value of 0.60. And the average of those all 5 samples is close to 0.61.

The number 0.606 is called the mean of SDOSP.

The standard deviation of all sample proportions

$$\sigma_{\hat{p}} = 0.049$$

This number 0.049 is called the standard deviation of SDOSP. It is also called the standard error (SE) of the mean.

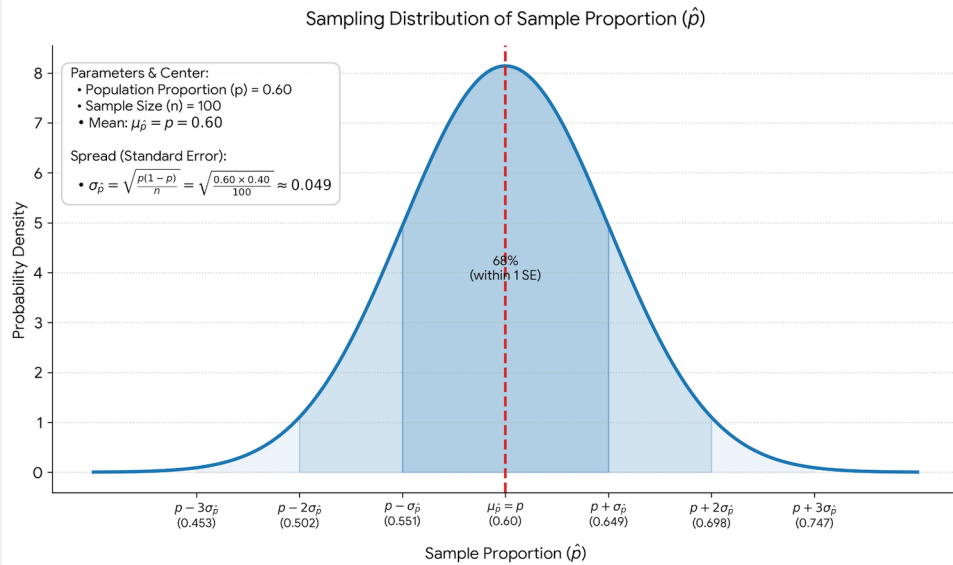


Figure 7.2.1: SDOSP is Normal

### Sampling Distribution of Sample Proportions and its Parameters

The **sampling distribution of the sample mean (SDOSP)** is the collection of all possible sample proportions  $\hat{p}$  taken from the same population using samples of the same size.

SDOSP is also denoted by  $\hat{p}$ . And  $\hat{p}$  is a random variable since it varies from sample to sample.

The parameters of SDOSP are  $\mu_{\hat{p}}$  and  $\sigma_{\hat{p}}$ .

### Underlying Population

The **underlying population** is the entire group of individuals, objects, or data from which samples are taken in a study. Unlike SDOSP, we don't discuss population distribution here in SDOSP because we are dealing with a categorical variable. But we will denote the population size by  $N$ .

### When Sampling Distribution of the Sample Proportion (SDOSP) Approximately Normal?

The sampling distribution of the sample proportion  $\hat{p}$  is approximately normal if the following conditions are satisfied:

1.  $np \geq 10$
2.  $n(1 - p) \geq 10$
3.  $N \geq 20n$

In other words, when these conditions are met, we can use the **Central Limit Theorem (CLT)** to model the sampling distribution of  $\hat{p}$  as approximately normal.

### Your Turn 7.2.1: Is SDOSP Normal?

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Definition: Mean and Standard Deviation of SDOSP

Suppose random samples of size  $n$  are drawn from a population in which the proportion with a characteristic of interest is  $p$ . The mean  $\mu_{\hat{p}}$  and standard deviation  $\sigma_{\hat{p}}$  of the SDOSP are given by the following formula if all three conditions that are mentioned above are met.

Mean of SDOSP

$$\mu_{\hat{p}} = p$$

Standard Deviation of SDOSP

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

### Your Turn 7.2.2: Population is Skewed

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 7.2.3: Describe SDOSP

Describe the sampling distribution of the sample proportion ( $\hat{p}$ ). Assume the size of the population is 10,000.

$$n = 100, p = 0.4$$

Here, describe means the following.

Verify whether the SDOSP is appropriately normal. Determine the mean and standard deviation of the sampling distribution of the sample proportion.

Answer available at the end of this section.

### Your Turn 7.2.3: Describe SDOSP

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Finding Probability Using SDOSP

### Example 7.2.4: Find Probability using SDOSP

Assume 85% of people actually show up for their flight on time. Because of this, airlines tend to overbook flights. A random sample of 236 booked passengers is taken, and whether or not they showed up on time is recorded for each of them.

1. Identify the individual, variable info from each individual, random variable  $X$ , and the random variable  $\hat{p}$  in the context of this problem.
2. Verify that the sampling distribution of the sample proportion  $\hat{p}$  is approximately normal.
3. What is the probability that at least 90% of the randomly selected booked passengers showed up on time?
4. Is 83% an unusually low percentage of booked passengers that show up on time for their flight in a sample of 236?

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

Answer 3 available at the end of this section.

### Your Turn 7.2.4: Find Probability using SDOSP

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example 7.2.5: Find Probability

An online retailer claims that 90% of all orders are shipped within 12 hours of being received. A consumer group placed 121 orders of different sizes and at different times of day; 102 orders were shipped within 12 hours.

1. Compute the sample proportion of items shipped within 12 hours.
2. Confirm that the sample is large enough to assume that the sample proportion is normally distributed. Use  $p = 0.90$ , corresponding to the assumption that the retailer's claim is valid.
3. Assuming the retailer's claim is true, find the probability that a sample of size 121 would produce a sample proportion so low as was observed in this sample.
4. Based on the answer to part (c), draw a conclusion about the retailer's claim.

Answer available at the end of this section.

### Your Turn 7.2.5: Find Probability

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Answers

### Example 7.2.3: Describe SDOSP

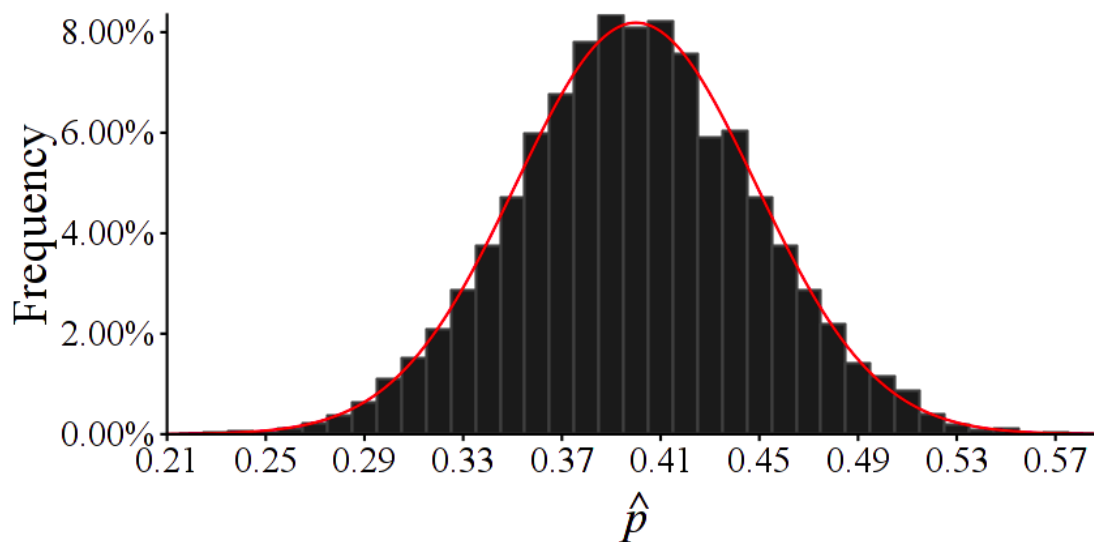


Figure 7.2.2: SDOSP is Normal

The sampling distribution of the sample proportion  $\hat{p}$  is approximately normal if the following conditions are satisfied:

1.  $np \geq 10$
2.  $n(1 - p) \geq 10$
3.  $N \geq 20n$

The first condition is satisfied, since

$$np = 100 \times 0.40 = 40 \geq 10$$

The second condition is satisfied, since

$$n(1 - p) = 100 \times (1 - 0.40) = 60 \geq 10$$

The third condition is satisfied, since

$$N \geq 20n \\ 10,000 \geq 20 \times 100 = 200$$

Since all the conditions are satisfied, SDOSP is approximately normal with

Mean of SDOSP

$$\mu_{\hat{p}} = p = 0.40$$

Standard deviation of SDOSP

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.40(1-0.40)}{100}} = 0.04899$$

### Example 7.2.4: Find Probability using SDOSP

**Individual:** a randomly selected booked passenger.

**Variable:** whether or not a randomly selected booked passenger shows up on time for their flight.

**Random variable  $X$ :** the number of 236 randomly selected booked passengers that show up on time for their flight.

**Random variable** Extra close brace or missing open brace : the percentage of 236 randomly selected booked passengers that show up on time for their flight

#### Example 7.2.4: Find Probability using SDOSP

The information given is that  $p = 0.85$ , hence  $1 - p = 0.15$ . First, we use the formulas to compute the mean and standard deviation of  $\hat{p}$ .

$$np = 236 \times 0.85 = 200.6$$

$$n(1-p) = 236 \times (1 - 0.85) = 35.4 \geq 10$$

$$N = 10,000 \geq 20 \times 236 = 4720$$

Now, the mean and standard deviation of SDOSP are given by

$$\mu_{\hat{p}} = p = 0.85 \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.85)(1-0.85)}{236}} = 0.02324$$

#### Example 7.2.4: Find Probability using SDOSP

$$P(\hat{p} \geq 0.90) = 0.0157$$

We compute probabilities using an **Online Statistics Calculator**

You can find  $P(\hat{p} \geq 0.90)$  using an online statistic calculator. Here are the steps.

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 0.85$  and  $\sigma = 0.02324$
3. Check two boxes  $P(x > a)$  and Calculate Area
4. Type  $P(0.90)$

Similarly, you can find  $P(\hat{p} \leq 0.83) = 0.1947$  using online statistics calculator. Since 0.1947 is greater than 0.05, 83% is not an unusually low percentage of booked passengers that show up on time for their flight in a sample of 236.

#### Example 7.2.5: Find Probability

1. The sample proportion is the number  $x$  of orders that are shipped within 12 hours divided by the number  $n$  of orders in the sample:

$$\hat{p} = \frac{x}{n} = \frac{102}{121} = 0.84$$

2. Since  $p = 0.90$ ,  $1 - p = 0.10$ , and  $n = 121$ ,

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.90)(0.10)}{121}} = 0.027$$

Because it is appropriate to use the normal distribution to compute probabilities related to the sample proportion  $\hat{p}$ .

3. Using the value of  $\hat{p}$  from part (a) and the computation in part (b),

$$\begin{aligned} P(\hat{P} \leq 0.84) &= P\left(Z \leq \frac{0.84 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) \\ &= P\left(Z \leq \frac{0.84 - 0.90}{0.027}\right) \\ &= P(Z \leq -2.20) \\ &= 0.0139 \end{aligned}$$

4. The computation shows that a random sample of size 121 has only about a 1.4% chance of producing a sample proportion as the one that was observed,  $\hat{p} = 0.84$ , when taken from a population in which the actual proportion is 0.90. This is so unlikely that it is reasonable to conclude that the actual value of  $p$  is less than the 90% claim.

This page titled [7.2: The Sampling Distribution of Sample Proportion](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by LibreTexts.

- [4.2: Surfaces and Area](#) by [Larry Green](#) is licensed [CC BY 4.0](#).
- [4.10: Stokes' Theorem](#) is licensed [CC BY 4.0](#).

## 8: ESTIMATING THE VALUE OF A PARAMETER

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.

8.1: Estimating Population Proportion

8.2: Estimating Population Mean

---

This page titled [8: Estimating the Value of a Parameter](#) 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: Estimating Population Proportion

### Introduction and Terminology for Estimating Population Proportion

Imagine trying to figure out exactly how many fish in a massive lake are a specific species, or what precise percentage of voters across the entire country plan to vote for a certain candidate. Unless you catch every single fish or poll every single citizen, knowing the absolute truth—the true **population proportion**—is practically impossible.

That is where estimation comes in. It bridges the gap between what we *can* measure (a small sample) and what we *want* to know (the whole population).

Estimating a population proportion is a statistical method used to estimate the percentage or fraction of an entire population that has a particular characteristic, based on information collected from a sample.

For example:

- A company estimates the proportion of customers who are satisfied with a product.
- A government estimates the unemployment rate.
- FSW wants to estimate the proportion of students on campus who smoke.
- A researcher wants to know what proportion of voters in a country support a particular candidate.

It would be difficult, expensive, and time-consuming to collect information from every individual in the population. Instead, the researcher surveys a smaller group, called a **sample**, and uses the sample results to estimate the true population proportion.

As you learned in the previous section:

$x$  = Number of individuals in the sample who have the characteristic

$n$  = Sample size

The **sample proportion** is calculated as:

$$\hat{p} = \frac{x}{n}$$

For example, suppose 120 out of 200 surveyed people prefer online shopping. Then:

$$\hat{p} = \frac{120}{200} = 0.60 = 60\%$$

Therefore, the sample proportion is 0.60, or 60%. We use this sample proportion to estimate the proportion of the **entire population** ( $p$ ) that prefers online shopping.

After estimating a population proportion using the sample proportion  $\hat{p}$ , we need to determine **how confident we are in that estimate**.

#### Point Estimate

The value of a sample statistic that estimates the corresponding population parameter. When we estimate the population proportion, the point estimate is  $\hat{p}$

$$\hat{p} = \frac{x}{n}$$

$x$  = Number of individuals in the sample who have the characteristic

$n$  = Sample size

## Notation and Formula Used When Estimating Population Proportion: ( $z$ )-Estimate

**Note:** The terminology used here is discussed somewhere below on this page or in the previous section.

Population proportion =  $p$  (It is unknown Parameter)

Sample proportion =  $\hat{p}$  (It is known Statistic)

The point estimate

$$\hat{p} = \frac{x}{n}$$

The point estimate is the center of the confidence interval. If we know LB and UB, we can find the point estimate  $\hat{p}$

$$\hat{p} = \frac{UB + LB}{2}$$

$z$  - critical value (CV) =  $z_{\frac{\alpha}{2}}$

Sample size =  $n$

Population size =  $N$

Number of individuals in the sample with the characteristic =  $x$

Confidence level =  $CL$

$$CL = 1 - \alpha$$

Total Area on the tails of the distribution =  $\alpha$

Area on the each tail of the distribution =  $\frac{\alpha}{2}$

$$\alpha = 1 - CL$$

The margin of error is the distance between the point estimate and one of the bounds. If we know LB and UB, we can find the margin of error  $E$ .

$$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

We can use the following formula to find  $\hat{p}$

$$E = \frac{UB - LB}{2}$$

Lower bound of the interval =  $LB$

$$LB = \hat{p} - E$$

Upper bound of the interval =  $UB$

$$UB = \hat{p} + E$$

## Confidence Interval

A confidence interval is an interval constructed around a point estimate that is likely to contain the corresponding population parameter. Each interval estimate is associated with a **confidence level**, which indicates how confident we are in the method used to estimate the population parameter.

We do not know the population proportion,  $p$ ; we have estimated it to lie between the upper and lower bounds.

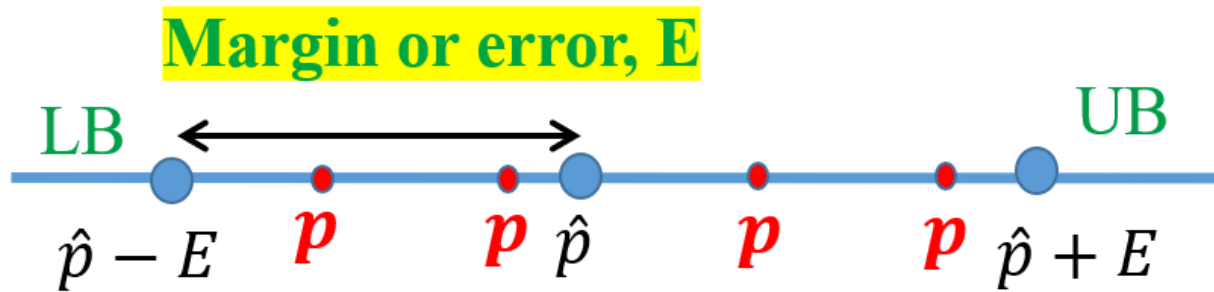


Figure  $\backslash(\backslash\text{PageIndex}\{1\}\backslash)$ : Graph of Confidence Interval

### Confidence Level (CL)

A **confidence level** (expressed as  $\backslash((1 - \alpha)\times 100\%\backslash)$ ) measures the long-run success rate of the statistical procedure used to construct a confidence interval. It represents the proportion of theoretical intervals that would contain the true population parameter if you were to repeat the sampling process indefinitely under identical conditions.

Think of trying to estimate an unknown population answer (like the true percentage of voters supporting a candidate) as casting a net into a lake to catch a target fish.

$\backslash(95\%\backslash)$  Confidence Level: Means that if you throw your net  $\backslash(100\backslash)$  separate times, your net will successfully catch the fish  $\backslash(95\backslash)$  out of  $\backslash(100\backslash)$  times.

### What do you need to construct a Confidence Interval for a Population Proportion?

To construct a confidence interval for the population proportion, we use the point estimate ( $\backslash(\hat{p}\backslash)$ ), which is the sample proportion in this case, and the margin of Error ( $\backslash(E\backslash)$ ). The point estimate is the center of the confidence interval.

After you have those two values, add  $\backslash(E\backslash)$  to  $\backslash(\hat{p}\backslash)$ , find the upper bound of the confidence interval, and subtract  $\backslash(E\backslash)$  from  $\backslash(\hat{p}\backslash)$  to find the lower bound of the confidence interval. See figure  $\backslash(\backslash\text{PageIndex}\{1\}\backslash)$  below.

### Different Forms of Confidence Intervals

We can write the confidence in three different forms

*Confidence in three different forms*

| Interval form                                                                     | Inequality form                   | Plus-Minus form                      |
|-----------------------------------------------------------------------------------|-----------------------------------|--------------------------------------|
| $\backslash((LB, UB)\backslash)$ or $\backslash(\hat{p}-E, \hat{p}+E)\backslash)$ | $\backslash(\hat{p}-E\backslash)$ | $\backslash(\hat{p}\pm E\backslash)$ |

The " $\backslash(\pm\backslash)$ " means:

Subtracting gives the lower bound of the confidence interval (LB).

Adding gives the upper bound of the confidence interval (UB).

### Example 1: Confidence Interval Formate

Suppose we know that a confidence interval is  $(0.6718, 0.6882)$ . Find the margin of error and point estimate, and also express the confidence interval in  $\hat{p} \pm E$  and  $LB$

Answer available at the end of this section.

### Your Turn 1: Confidence Interval Formate

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Criteria to Construct a Confidence Interval for a Population Proportion

What are the conditions that must be met to construct a confidence interval for a population proportion?

1. The sample must be a simple random sample
2.  $n\hat{p} \geq 10$  and  $n(1-\hat{p}) \geq 10$
3.  $N \geq 20n$

### Interpretation of Confidence Interval

"We are [Confidence Level]% confident that the true population proportion of [context of the problem] is between [lower bound] and [upper bound]." Suppose confidence level is 95%

1. There is a 95% chance that  $\hat{p} - E$  or
2. A 95% level of confidence implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, we will expect 95 of the intervals to contain the parameter and 5 not to include the parameter.

### Example 2: Interpretation of Confidence Interval

Gallup conducted a poll and found a 95% confidence interval of 0.56

Answer available at the end of this section.

### Your Turn 2: Interpretation of Confidence Interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



# Critical Values

It is a boundary value that determines how many standard errors you need to move away from the sample statistic to achieve your desired level of confidence.

Think of it as a multiplier that adjusts the width of your interval based on how certain you want to be.

## Why do you need to find the Critical value (CV) or $(z_{\frac{\alpha}{2}})$

The **critical value (CV)** is a factor used to calculate the margin of error in a confidence interval. If you are given a confidence level, you can find the critical values.

The critical value is derived from a theoretical sampling distribution (usually a standard normal  $(z)$ -distribution or a  $(t)$ -distribution (section (8.2)).

## Graphical Representation of Critical Value

When you choose a confidence level (e.g., (95%)), you are deciding to capture the middle (95%) of the distribution.

The critical values are the points on the horizontal axis that cut off the remaining "tails" (the remaining (5%)), split into (2.5%) on the left and (2.5%) on the right.

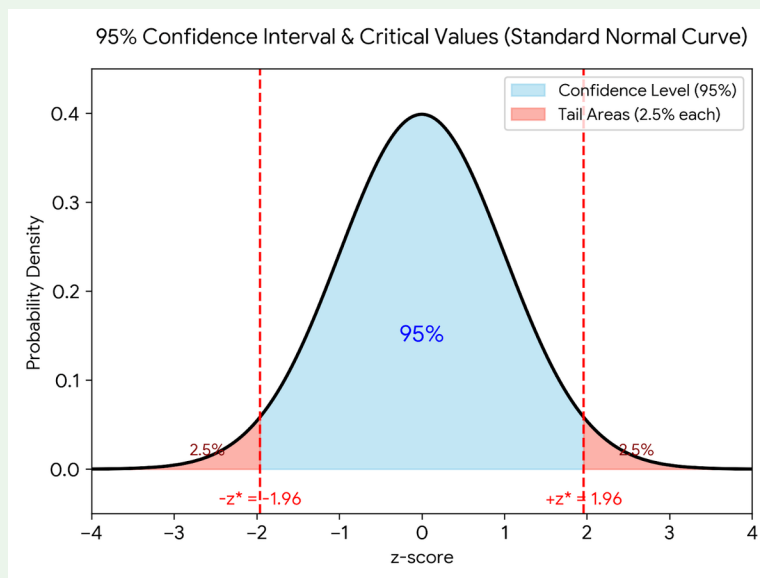


Figure (PageIndex{2}): Representation of Critical Values Graphically

In the above example, there are two critical values  $(-1.96)$  and  $(1.96)$  for (95% ) confidence level.

We call left critical value  $((-z_{\frac{\alpha}{2}}=-1.96))$  and right Critical value  $((z_{\frac{\alpha}{2}}=1.96))$ . Alpha  $(\alpha=0.05)$  is the total area in the tail, and  $(\frac{\alpha}{2}=0.025)$  is the area on each side of the tail. The total between two critical values is (0.95).

## How to Find the Critical Value if CL is Given?

We can find the critical value using a standard normal table or a statistical software package. The video attached at the bottom of this page demonstrates how to find the critical value for a given CL. It is strongly recommended that you draw and label the standard normal curve and mark the critical values on the graph before proceeding with your calculations.

### Your Turn $\backslash(\backslash\text{PageIndex}\{3\}\backslash)$ : Find Critical Value

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Constructing a Confidence Interval for a Population Proportion

### Steps to Construct a Confidence Interval for the Population Proportion

Step 1: Identify the Parameter and Sample Statistics by carefully reading the question.

Step 2: Verify the Conditions

Step 3: Perform the Calculations: Find the Critical value and margin of error.

Step 4: Find the Interval using  $\backslash(\text{Point Estimate}) \pm \text{Margin of Error}\backslash)$

Step 4: Interpret the Interval (Conclude)

**Note:** You can find a confidence interval directly using technology. Please, watch the attached video.

### Example $\backslash(\backslash\text{PageIndex}\{4\}\backslash)$ : Confidence Interval for the Population Proportion

A sociologist wants to construct a  $\backslash(99\%\backslash)$  confidence interval for the true proportion of children aged  $\backslash(8-10\backslash)$  living in New York who own a cell phone. Suppose a simple random sample of size  $\backslash(800\backslash)$  is taken and it is found that  $\backslash(480\backslash)$  children aged  $\backslash(8-10\backslash)$  have a cell phone in New York.

1. What is the sample size, and the number of individuals with given characteristics?
2. State the parameter in words.
3. Find the sample statistic (point estimate).
4. State and check the assumptions for a confidence interval.
5. Find the critical value and the margin of error.
6. Find the confidence interval for the proportion of children aged  $\backslash(8-10\backslash)$  who have a cell phone in New York.
7. Interpret the confidence interval.

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

Answer 3 available at the end of this section.

Answer 4 available at the end of this section.

Answer 5 available at the end of this section.

Answer 6 available at the end of this section.

Answer 7 available at the end of this section.

### Your Turn $\backslash(\backslash\text{PageIndex}\{4\}\backslash)$ : Construct Confidence Interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example $\backslash(\backslash\text{PageIndex}\{5\}\backslash)$ : Construct Confidence Interval

City officials want to estimate the proportion of residents who regularly recycle. A random sample of  $\backslash(284\backslash)$  residents found that  $\backslash(37\backslash)$  reported recycling regularly.

1. What is the best point estimate for the percentage of all residents in the city who recycle regularly? (Enter your answer as a percent rounded to two decimal places.)
2. Construct a  $\backslash(90\%\backslash)$  confidence interval estimate for the percentage of all residents who recycle regularly. (Round both bounds to two decimal places and expressed in percentage.)
3. What is the margin of error for this confidence interval? Use the confidence interval from above to find the margin of error.
4. Write the confidence interval in the form  $\backslash(\hat{p} \pm E\backslash)$ .
5. Interpret the confidence interval.

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

Answer 3 available at the end of this section.

Answer 4 available at the end of this section.

Answer 5 available at the end of this section.

### Your Turn $\backslash(\backslash\text{PageIndex}\{5\}\backslash)$ : Find Confidence Interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Sample Size Consideration

The confidence interval formula for estimating a population proportion  $\backslash(p\backslash)$  is  $\backslash(\hat{p} \pm E\backslash)$ , where

$$\backslash[E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\backslash]$$

The number  $\backslash(z_{\alpha/2}\backslash)$  is determined by the desired level of confidence. To say that we wish to estimate the population proportion to within a certain number of percentage points means that we want the margin of error  $\backslash(E\backslash)$  to be no larger than that number (expressed as a proportion). Thus, we obtain the minimum sample size needed by solving the displayed equation for  $\backslash(n\backslash)$ .

## Minimum Sample Size for Estimating a Population Proportion

The estimated minimum sample size  $n$  needed to estimate a population proportion  $p$  to within  $E$  at  $(1-\alpha)\%$  confidence is

If previous estimate  $\hat{p}$  is known

$$n = \frac{\hat{p}(1-\hat{p})}{\left(\frac{z_{\alpha/2}}{E}\right)^2}$$

If previous estimate  $\hat{p}$  is not known, use  $\hat{p}=0.50$  in above formula.

There is a dilemma here: the formula for estimating how large a sample to take contains the number  $\hat{p}$ , which we know only after we have taken the sample. There are two ways out of this dilemma. Typically, the researcher will have some idea as to the value of the population proportion  $p$ , hence of what the sample proportion  $\hat{p}$  is likely to be. For example, if last month  $(37\%)$  of all voters thought that state taxes are too high, then it is likely that the proportion with that opinion this month will not be dramatically different, and we would use the value  $(0.37)$  for  $\hat{p}$  in the formula.

The second approach to resolving the dilemma is simply to replace  $\hat{p}$  in the formula by  $(0.5)$ . This is because if  $\hat{p}$  is large, then  $(1-\hat{p})$  is small, and vice versa, which limits their product to a maximum value of  $(0.25)$ , which occurs when  $\hat{p} = 0.5$ . This is called the **most conservative estimate**, since it gives the largest possible estimate of  $n$ .

### Example 6: Find the Sample Size

Find the necessary minimum sample size to construct a  $(98\%)$  confidence interval for  $p$  with a margin of error  $E=5\%$ .

- assuming that no prior knowledge about  $p$  is available; and
- assuming that prior studies suggest that  $p$  is about  $(0.1)$ .

Answer 1 available at the end of this section.

Answer 2 available at the end of this section.

Answer available at the end of this section.

### Your Turn 6: Find the Sample Size

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Answers

### Example 1: Confidence Interval Format

Margin of Error:

$$E = \frac{(0.6882 - 0.6718)}{2} = 0.0082$$

Point Estimate:

$$\hat{p} = \frac{(0.6718 + 0.6882)}{2} = 0.68$$

Interval in Plus-minus form

$\hat{p} = 0.68$

Interval in Inequality form:

$[0.6718$

#### Example 1: Interpretation of Confidence Interval

With 95% confidence, it can be said that the true proportion of Americans who believe it is the government's responsibility for health care is between 56% and 61%.

If we constructed 100 of these confidence intervals for the same population with the same sample size, we would expect 95 of them to contain the true population proportion, and 5 not to.

#### Example 4: Confidence Interval for the Population Proportion

$x = 480$  and  $n = 800$

#### Example 4: Confidence Interval for the Population Proportion

The proportion of all children in New York who own a cell phone.

#### Example 4: Confidence Interval for the Population Proportion

Point estimate  $\hat{p}$

$$\hat{p} = \frac{x}{n} = \frac{480}{800} = 0.6$$

#### Example 4: Confidence Interval for the Population Proportion

The sample is assumed to be random, and

$$n\hat{p} = 800 \cdot 0.6 = 480 \geq 10$$

$$n(1-\hat{p}) = 800 \cdot (1-0.6) = 192 \geq 10$$

$20n = 20 \cdot 800 = 16,000 \leq \text{all } (8-10) \text{ years old living in New York}$

So all conditions are met

#### Example 4: Confidence Interval for the Population Proportion

For 99% confidence interval:  $CL = 0.99$ ,  $\alpha = 1 - CL = 1 - 0.99 = 0.01$

Left critical value:  $-z_{\frac{\alpha}{2}} = -2.576$

Right critical value:  $z_{\frac{\alpha}{2}} = 2.576$

(Note: Watch the attached video to show how to find the critical values using the online statistics Calculator. We also did find the critical values for the given confidence level in Your Turn 3).

Margin of Error (E)

$$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 2.576 \sqrt{\frac{0.6(1-0.6)}{800}}$$

$$\approx 0.0446$$

#### Example 4: Confidence Interval for the Population Proportion

Confidence Interval

$$\hat{p} \pm E$$

$$0.6 \pm 0.0446$$

$$\text{Interval form: } (0.5554, 0.6446)$$

$$\text{Inequality form: } 0.5554 < p < 0.6446$$

$$0.5554 < p < 0.6446$$

#### Example 4: Confidence Interval for the Population Proportion

We 99% are confident that the true proportion of children aged 8-10 living in New York who own a cell phone is between 55.54% and 64.46%

#### Example 5: Construct Confidence Interval

Here  $n = 284$  and  $x = 37$ . Point estimate here is sample proportion  $\hat{p}$

$$\hat{p} = \frac{x}{n} = \frac{37}{284} = 0.1303 = 13.03\%$$

#### Example 5: Construct Confidence Interval

In the online statistics calculator. Do the following steps

1. Put the cursor under **Estimates**

2. Click **z Confidence Interval for a proportion**

3. Put  $n = 284$  and  $x = 37$ , and  $c = 0.90$

4. Click **Calculate**

You will get  $(0.0974, 0.1631)$ . As a percentage confidence interval, which will be

$(9.74\%, 16.31\%)$

#### Example 5: Construct Confidence Interval

$$E = \frac{UB-LB}{2} = \frac{16.31-9.74}{2} = 3.29\%$$

#### Example 5: Construct Confidence Interval

$$\hat{p} \pm E = 13.03\% \pm 3.29\%$$

#### Example 5: Construct Confidence Interval

We are 90% confident that the interval from 9.74% to 16.32% actually contains the true value of the population proportion of all residents in the city who recycle regularly

**Example 1: Find the Sample Size**

Confidence level 98% means that  $\alpha = 1 - 0.98 = 0.02$  so  $\alpha/2 = 0.01$ .

Using an online statistics calculator, we can find the critical value

$$z_{\frac{\alpha}{2}} = 2.326$$

Since there is no prior knowledge of  $p$ , we make the most conservative estimate that  $\hat{p} = 0.5$ . Then

$$\begin{aligned} n &= \hat{p}(1 - \hat{p}) \left( \frac{z_{\alpha/2}}{E} \right)^2 \\ &= 0.50(1 - 0.50) \left( \frac{2.326}{0.05} \right)^2 \\ &= 541.0276 \end{aligned}$$

Which we round UP to 542.

**Example 2: Find the Sample Size**

Since  $p \approx 0.10$ , we estimate  $\hat{p}$  by 0.10, and obtain

$$\begin{aligned} n &= \hat{p}(1 - \hat{p}) \left( \frac{z_{\alpha/2}}{E} \right)^2 \\ &= 0.10(1 - 0.10) \left( \frac{2.326}{0.05} \right)^2 \\ &= 194.769936 \end{aligned}$$

Which we round UP to 195.

**Example 3: Find the Sample Size**

- [Green's Theorem](#) by [Larry Green](#) is licensed [CC BY 4.0](#).

## 8.2: Estimating Population Mean

---

### Introduction and Terminology for Estimating Population Mean

Suppose the university has 20,000 graduating seniors. And you want to know the average debt of all 20,000 graduating seniors. You cannot ask all of them, so you randomly survey 100 students and calculate their average debt.

Your sample mean  $\bar{x}$  comes out to exactly \$28,500.

If you stop there and tell the university administration that the average debt of our graduates is \$28,500, you imply absolute certainty. But what if you had picked a different 100 students? You might have gotten \$27,200 or \$29,900. Because of sampling variability, it is highly unlikely that your specific sample average perfectly matches the true average of all 20,000 students  $\mu$ .

### The Solution: The Confidence Interval

Instead of risking a single, likely imperfect guess, you calculate a 95% **confidence interval** using your sample data. Your calculation accounts for the sample size 100 and the variability in debt amounts from student to student (the standard deviation).

Your final result might look like this:

\$28,500  $\pm$  \$1,200, which gives a range of \$27,300 to \$29,700

### Why This Is Much More Useful

1. It admits sampling error exists: By providing a range \$27,300 to \$29,700, you openly acknowledge that a sample is just a snapshot, not the whole picture.
2. It measures precision: If your interval was incredibly wide—say, \$15,000 to \$42,000—the administration would know the data is too erratic to trust for budgeting. A tighter interval (\$27,300 to \$29,700) proves you have a highly precise estimate.
3. It provides a safety net for decision-making: If the university wants to launch a financial aid initiative for anyone with above-average debt, planning within a realistic range ensures they don't underfund the program based on a single, fluky sample mean.

In short, we construct confidence intervals because samples give us clues, not absolute facts. The interval turns a single, uncertain guess into a reliable, statistically backed range.

### Comparing the $z$ and $t$ Approaches

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  $\sigma$  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  $\sigma$  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 current practice is to use the Student's  $t$ -distribution whenever  $s$  is used as an estimate of  $\sigma$ . If you draw a simple random sample of size  $n$  from a population that has an approximately normal distribution with mean  $\mu$  and unknown population standard deviation  $\sigma$ , and calculate the  $t$ -score

$$t = \frac{\bar{x} - \mu}{\left( \frac{s}{\sqrt{n}} \right)}$$

then the  $t$ -scores follow a Student's  $t$ -distribution with  $n-1$  degrees of freedom. The  $t$ -score has the same interpretation as the  $z$ -score. It measures how far  $\bar{x}$  is from its mean  $\mu$ . 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).

The table below highlights how **knowing** or **not knowing** the population standard deviation alters your statistical approach:

*z-interval vs t-interval*

| Condition            | $\sigma$ is known                                  | $\sigma$ is Unknown                                 |
|----------------------|----------------------------------------------------|-----------------------------------------------------|
| Distribution Used    | Standard Normal $z$ -distribution                  | Student's $t$ -distribution                         |
| Critical Value       | $\pm z_{\frac{\alpha}{2}}$                         | $\pm t_{\frac{\alpha}{2}}$                          |
| Name of the Interval | $z$ -interval for population mean ( $z$ -estimate) | $t$ - interval for population mean ( $t$ -estimate) |

### Notation and Formula Used When Estimating Population Proportion: ( $z$ -Estimate or $t$ -Estimate)

Population mean =  $\mu$  (It is unknown parameter)

Sample mean =  $\bar{x}$  (It is known statistic also called point estimate)

Population standard deviation =  $\sigma$  (May be know or unknown Parameter)

Sample standard deviation =  $s_x$  or  $s$

The point estimate  $\bar{x}$  is the center of the confidence interval. We can find it if the LB and UB are given using the following formula.

$$\hat{x} = \frac{UB + LB}{2}$$

$z$  - critical value =  $z_{\frac{\alpha}{2}}$  (For  $z$ -distribution when  $\sigma$  is known)

$t$  - critical value =  $t_{\frac{\alpha}{2}}$  (For  $t$ -distribution when  $\sigma$  is unknown)

Sample size =  $n$

Population size =  $N$

Degree of freedom =  $df = n - 1$

Confidence level =  $CL$  or  $C$

$$CL = 1 - \alpha$$

Total Area on the tail of the distribution =  $\alpha$

$$\alpha = 1 - CL$$

Total Area on the each tail =  $\frac{\alpha}{2}$

Margin of error =  $E$

If  $\sigma$  is known

$$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

If  $\sigma$  is unknown

$$E = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

The margin of error  $E$  is the distance between the point estimate and one of the bounds. We can find it if the LB and UB are given as follows.

$$E = \frac{UB - LB}{2}$$

Lower bound of the confidence interval =  $LB$

$$LB = \bar{x} - E$$

Upper bound of the confidence interval =  $UB$

$$UB = \bar{x} + E$$

### What do you need to construct a Confidence Interval for a Population Mean?

To construct a confidence interval for the population proportion, we use the point estimate, which is the sample mean in this case, and the margin of Error. The point estimate is the center of the confidence interval.

After you have those two values, add  $E$  to  $\bar{x}$ , find the upper bound of the confidence interval, and subtract  $E$  from  $\bar{x}$  to find the lower bound of the confidence interval. See figure 8.2.1 below.

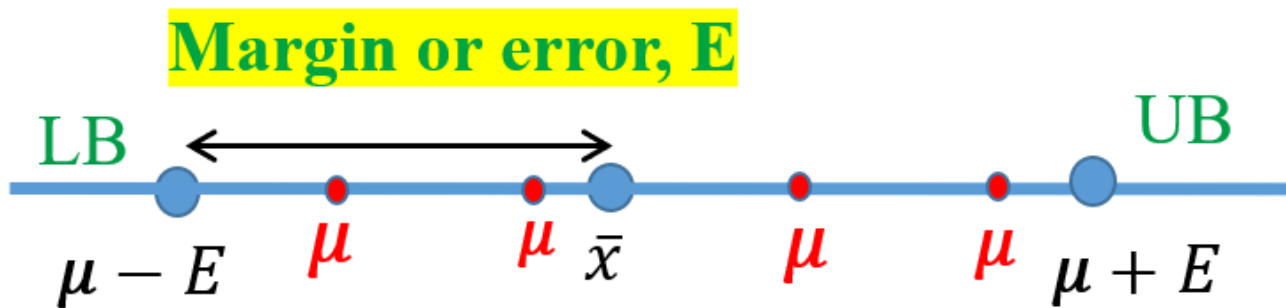


Figure 8.2.1: Graphical Representation of Confidence Interval for  $\mu$

We do not know the population means; we have estimated it to lie between the upper and lower bounds.

### Different forms of Confidence Interval

We can write the confidence in three different forms

### Different forms of Confidence Interval

| Interval form                              | Inequality form                   | Plus-Minus form |
|--------------------------------------------|-----------------------------------|-----------------|
| $(LB, UB)$ or $(\bar{x} - E, \bar{x} + E)$ | $\bar{x} - E < \mu < \bar{x} + E$ | $\bar{x} \pm E$ |

The “ $\pm$ ” means:

Subtracting gives the lower bound of the confidence interval (LB).

Adding gives the upper bound of the confidence interval (UB)

### Example 8.2.1: Different forms of Confidence Interval

Suppose we know that a confidence interval is (67.18, 68.82). Find the margin of error and point estimate, and also express the confidence interval in  $\bar{x} \pm E$  and  $LB < \mu$

Answer available at the end of this section.

### Your Turn 8.2.1: Different forms of Confidence Interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Criteria to Construct a Confidence Interval for a Population Proportion

What conditions must be checked to construct a confidence interval for a population proportion?

1. The sample must be a simple random sample
2. Either the sample size  $n \geq 30$  or the population needs to be normally distributed.
3. Check if the  $\sigma$  is known or not.

Note: If population standard deviation  $\sigma$  is known, we need to use  $z$ -estimate, and if population standard deviation  $\sigma$  is unknown and sample standard deviation  $s$  is known, we use  $t$ -estimate.

### Interpretation of Confidence Interval

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  $CL\%$  confidence that the true population mean (include the context of the problem) is between  $LB$  and  $UB$  (include appropriate units)."

1. There is a 95% chance that  $\bar{x} - E < \mu < \bar{x} + E$  contains the true mean  $\mu$  which is unknown." **Or**
2. A 95% level of confidence implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, we will expect 95 of the intervals to contain the parameter and 5 not to include the parameter.

### Example 8.2.2: Interpretation of Confidence Interval

A researcher constructs a 95% **confidence interval** for the mean number of hours that college students study each week. The confidence interval is:  $12.5 < \mu < 16.3$ .

Interpret 95% confidence interval.

*Answer available at the end of this section.*

### Your Turn 8.2.2: Interpretation of Confidence Interval

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Critical Values for $t$

The **critical value (CV)** is a factor used to calculate the margin of error in a confidence interval. It is a boundary value that determines how many standard errors you need to move away from the sample statistic to achieve your desired level of confidence.

### What do you need to find the Critical Values for $t$ ?

If you are given a confidence level and degrees of freedom, you can find the critical values for  $t$  distribution. Think of it as a multiplier that adjusts the width of your interval based on how certain you want to be.

### Graphical Representation of Critical Value of $t$

When you choose a confidence level (e.g., 95%), you are deciding to capture the middle 95% of the  $t$  - distribution. The critical values are the coordinates on the horizontal axis that cut off the remaining "tails" (the remaining 5%), split into 2.5% on the left and 2.5% on the right.

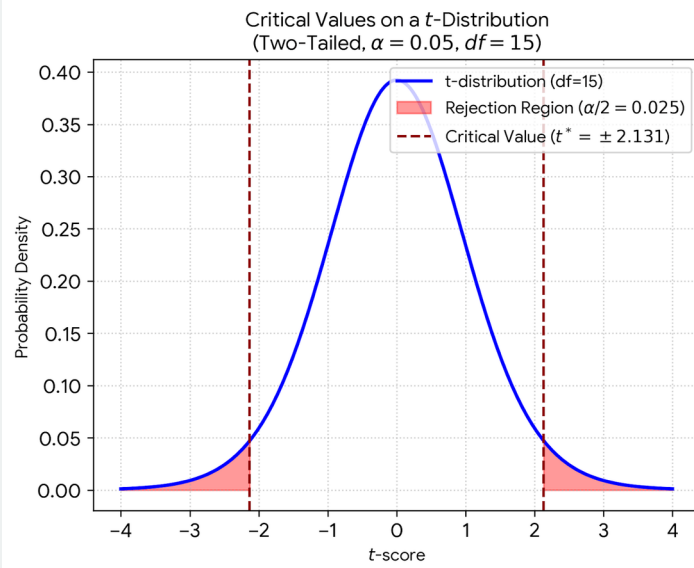


Figure 8.2.2: Representation of Critical Value for  $t$  Graphically

In the above example, there are two critical values for  $t$ ,  $-2.131$  and  $2.131$  for 95% confidence level.

We call left critical value ( $-t_{\alpha/2} = -2.131$ ) and right Critical value ( $t_{\alpha/2} = 2.131$ ). Alpha  $\alpha$  is the total area in the tail, and  $\frac{\alpha}{2}$  is the area on each side of the tail.

### How to find the Critical Value?

We can find the critical value using a statistical table or a statistical software package. The video attached at the bottom of this page demonstrates how to find the critical value. It is strongly recommended that you draw and label the standard normal curve and mark the critical value on the graph before proceeding with your calculations.

### Your Turn 8.2.3: Find Critical Value for $t$

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Constructing a Confidence Interval for a Population Mean

### Steps to Construction Confidence Interval For Population Mean

- Step 1: Identify the parameter and sample statistic by carefully reading the question. And also identify whether you are using  $z$ -estimate or  $t$ -estimate.
- Step 2: Verify the conditions
- Step 3: Perform the calculations: Find the critical value and margin of error.
- Step 4: Find the interval using: Point Estimate  $\pm$  Margin of Error

Step 4: Interpret the interval (conclusion)

**Note:** You can find a confidence interval directly using technology. Please, watch the attached video.

#### Example 8.2.4: Confidence Interval for the Population mean $\sigma$ is Known

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 yields 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.

1. What is the sample size and other information given?
2. State the parameter in words.
3. Find the sample statistic (point estimate).
4. State and check the assumptions to construct a confidence interval for a population mean and decide whether it is  $z$ -estimate or  $t$ -estimate.
5. Find the critical values.
6. Find the margin of error.
7. Find the confidence interval for the proportion of children aged 8 – 10 who have a cell phone in New York.
8. Interpret the confidence interval.

*Answer 1 available at the end of this section.*

#### Your Turn 8.2.4: Construct Confidence Interval ( $\sigma$ ) is Known

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



The following example shows how to find the confidence interval if ( $\sigma$ ) is unknown and data is given.

#### Example 8.2.5: Confidence Interval for the Population mean $\sigma$ is Unknown

A researcher wants to estimate the mean weight of all pumpkins grown on a farm. Assume that the weight of all pumpkins is normally distributed. A farmer selected a random sample of 25 pumpkins and found its mean to be 18.4 pounds and the sample standard deviation to be 3.5 pounds. Construct a confidence interval for a 95% confidence level.

*Answer available at the end of this section.*

### Your Turn 8.2.5: Find Confidence Interval ( $\sigma$ ) is unknown

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Sample Size Consideration

### Minimum Sample Size for Estimating a Population Mean

The estimated minimum sample size  $n$  needed to estimate a population mean  $\mu$  to within  $E$  at  $(1 - \alpha)\% \times 100$  confidence is

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

Round UP to the nearest whole number.

### Example 8.2.6: Find Sample Size

An economist wishes to estimate, with a 95% confidence interval, the yearly income of welders with at least five years of experience to within \$1,000. He estimates that the income range is no more than \$24,000, so using the Empirical Rule, he estimates the population standard deviation to be about one-sixth as large, or about \$4,000. Find the estimated minimum sample size required.

*Answer available at the end of this section.*

### Your Turn 8.2.6: Find the Sample Size

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Changing the Confidence Level or Sample Size

### Example 8.2.7: Effect of $CL$ and $n$ in Confidence Interval

Refer the Example 8.2.4. We found 90% confidence level for  $n = 36$ . Now

- Find a 95% confidence interval for the true (population) mean of the statistics exam score for  $n = 36$ .
- Find a 90% confidence interval for the true (population) mean of the statistics exam score for  $n = 100$  instead of  $n = 36$ ?

*Answer available at the end of this section.*

### Your Turn 8.2.7: Effect of $CL$ and $n$ in Confidence Interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



### Example 8.2.8: 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.

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

1. Find the point estimate

Answer available at the end of this section.

### Your Turn 8.2.8: Construct Confidence Interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Reference

1. "America's Best Small Companies." Forbes, 2013. Available online at [Froeb's](http://www.forbes.com) [www.forbes.com] (accessed July 2, 2013).
2. "Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012." Federal Election Commission. Available online at [www.fec.gov/data/index.jsp](http://www.fec.gov/data/index.jsp) (accessed July 2, 2013).
3. "Human Toxome Project: Mapping the Pollution in People." Environmental Working Group. Available online at [www.ewg.org/sites/humantoxome...tero%2Fnewborn](http://www.ewg.org/sites/humantoxome...tero%2Fnewborn) (accessed July 2, 2013).
4. "Metadata Description of Leadership PAC List." Federal Election Commission. Available online at [www.fec.gov/finance/disclosur...pPacList.shtml](http://www.fec.gov/finance/disclosur...pPacList.shtml) (accessed July 2, 2013).

## Answers

### Example 8.2.1: Different forms of Confidence Interval

Margin of Error:

$$E = \frac{(68.82 - 67.18)}{2} = 0.82$$

Point Estimate:

$$\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$$

Different forms of confidence intervals are

$$\bar{x} \pm E = 68 \pm 0.82$$

and

$$67.18 < \mu < 68.82$$

### Example 8.2.2: Interpretation of Confidence Interval

We are 95% **confident** that the true population mean number of hours that all college students study per week is between 12.5 hours and 16.3 hours.

If we constructed 100 of these confidence intervals for the same population with the same sample size, we would expect 90 of them to contain the true population mean exam score, and 10 not to.

### Example 8.2.4: Confidence Interval for the Population mean $\sigma$ is Known

$\bar{x} = 68$ ,  $\sigma = 3$ , and  $n = 36$ . The normal population is also given.

Population mean ( $\mu$ ) (the mean score on all exams).

Point estimate  $\bar{x}$  here is sample mean of 36 scores.

$$\bar{x} = 68$$

The sample is assumed to be random, and the population distribution is unknown, but  $n = 36 \geq 30$ , and we know the population standard deviation. We use  $z$ -estimate here.

The confidence level is 90%,  $CL = 0.90$

$$\alpha = 1 - CL = 1 - 0.90 = 0.10$$

Left critical value:  $-1.645$

Right critical value:  $1.645$

$$z_{\frac{\alpha}{2}} = \pm 1.645$$

**Note:** Watch the attached video to learn how to find the critical value using a statistics online calculator.

Margin of Error

$$\begin{aligned} E &= z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \\ &= 1.645 \cdot \frac{3}{\sqrt{36}} \\ &= 0.8225 \end{aligned}$$

$$LB = \bar{x} - E = 68 - 0.8225 = 67.18$$

$$UB = \bar{x} + E = 68 + 0.8225 = 68.82$$

Different forms of confidence intervals

Plus minus form

$$68 \pm 0.8225$$

Interval form

$$(67.18, 68.82)$$

Inequality form

$$67.18 < \mu < 68.82$$

We estimate with 90% confidence that the true mean exam score for all statistics students is between 67.18 and 68.82.

Ninety percent of all confidence intervals constructed in this way contain the true mean of the statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population means exam score.

### Example 8.2.5: Confidence Interval for the Population mean $\sigma$ is Unknown

$\bar{x} = 18.4$ ,  $s = 3.5$ , and  $n = 25$

The unknown parameter is the mean weight of all pumpkins

Here, point estimate is  $\bar{x} = 18.4$  sample mean.

The sample is assumed to be random, and the population is normal, but we do not know the population standard deviation. So we are using  $t$ -estimate for population mean.

The confidence level is 95%,  $CL = 0.95$ .

Degree of freedom:  $df = n - 1 = 25 - 1 = 24$

$$\alpha = 1 - CL = 1 - 0.95 = 0.05$$

Left critical value for  $t$ :  $-2.064$  and Right critical value for  $t$ :  $2.064$

$$t_{\frac{\alpha}{2}} = \pm 2.064$$

NOTE: Watch the attached video to learn how to find the critical value using  $df$  and  $CL$ .

Margin of Error

$$\begin{aligned} E &= t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \\ &= 2.064 \cdot \frac{3.5}{\sqrt{25}} \\ &= 1.44 \end{aligned}$$

Confidence Interval

$$LB = \bar{x} - E = 18.4 - 1.44 = 16.96$$

$$UB = \bar{x} + E = 18.4 + 1.44 = 19.84$$

Interpretation

We estimate with 95% confidence that the true mean weight of all pumpkins grown on the farm is between 16.96 pounds and 19.84 pounds.

If we constructed 100 of these confidence intervals, we would expect 95 of them to contain the true mean weight of all pumpkins, and 5 do not.

#### Example 8.2.6: Find Sample Size

Confidence level 95% means that  $\alpha = 1 - 0.95 = 0.05$  so  $\frac{\alpha}{2} = 0.025$ .

We can find the critical value using an online statistic calculator (Refer to sec 8.1)

$$z_{\frac{\alpha}{2}} = 1.960$$

To say that the estimate is to be "to within \$1,000" means that  $E = 1000$ . Thus

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 \sigma^2}{E^2} \\ &= \frac{(1.960)^2 (4000)^2}{(1000)^2} \\ &= 61.4656 \end{aligned}$$

Which we round up to 62.

#### Example 8.2.7: Effect of $CL$ and $n$ in Confidence Interval

For  $CL = 90\%$  and  $n = 36$

The 90% confidence interval is (67.18, 68.82) and  $E = 0.82$ .

Now for  $CL = 95\%$  and  $n = 36$  (Just increase  $CL$ )

In the online statistics calculator. Do the following steps

1. Put the cursor under **Estimates**
2. Click **the Confidence Interval for the mean**
3. Put  $\bar{x} = 68$ , ( $\sigma=3$ ),  $n = 36$ , and  $c = 0.95$
4. Check box "Input form Stats"
4. Click **Calculate**

You will get (67.02, 68.98)

Margin of Error:

$$E = \frac{(68.98 - 67.02)}{2} = 0.98$$

The 95% confidence interval  $n = 36$  (67.02, 68.98) and  $E = 0.98$ .

The 95% confidence interval is wider (Bigger margin of error).

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.

Increasing the confidence level increases the error bound, making the confidence interval wider.

Let's see what happens to the confidence interval if we change the sample size.

Now for  $CL = 90\%$  and  $n = 100$  (Just increase  $n$ )

$$\text{When } n = 100 : E = \left( z_{\frac{\alpha}{2}} \right) \left( \frac{\sigma}{\sqrt{n}} \right) = (1.645) \left( \frac{3}{\sqrt{100}} \right) = 0.4935.$$

In the online statistics calculator. Do the following steps

1. Put the cursor under **Estimates**
2. Click **the Confidence Interval for the mean**
3. Put  $\bar{x} = 68$ , ( $\sigma=3$ ),  $n = 100$ , and  $c = 0.90$
4. Check box "Input form Stats"
4. Click **Calculate**

You will get (67.51, 68.49)

Margin of Error:

$$E = \frac{(68.49 - 67.51)}{2} = 0.49$$

The 90% confidence interval with  $n = 100$  is (67.51, 68.49) and  $E = 0.49$ .

If we **increase** the sample size  $n$  to 100, we **decrease** the error bound, making the confidence interval narrower.

#### Example 8.2.8: Acupuncture

1. Click List Editor.
2. Put the given data in L1 (there are 15 data)
3. Put the cursor over Estimate and Select T Interval for mean (Since  $\sigma$  is known)
4. Check the box "Input form Data"
5. List should be L1 (since you put the data in L1) and type  $c = 0.95$
6. Press Calculate

You can see the confidence interval 7.3, 9.2).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

---

This page titled [8.2: Estimating Population Mean](#) is shared under a license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12: Supplementary notes on matrices and linear systems](#) by [Isaiah Lankham](#), [Bruno Nachtergaele](#), and [Anne Schilling](#) has no license indicated.
- [Green's Theorem](#) by [Larry Green](#) is licensed [CC BY 4.0](#).

## 9: HYPOTHESIS TESTING

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.

9.1: Language of Hypothesis Testing

9.2: Test for the Population Proportion

9.3: Test for the Population Mean

---

This page titled [9: 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.

## 9.1: Language of Hypothesis Testing

### Hypothesis

The actual test begins by considering two **hypotheses**. They are called the **null** and **alternative hypotheses**. These hypotheses contain opposing viewpoints.

#### Hypothesis

In hypothesis testing, a **hypothesis** is a quantitative claim or statement about an unknown population parameter (such as a mean  $\mu$  or proportion  $p$ ) that is evaluated using sample data.

#### Null Hypothesis $H_0$

A null hypothesis is an assumption that nothing has changed or that there is no effect. It is a hypothesis to be tested. A null hypothesis (usually written as  $H_0$ ) is the default claim that there is no effect, no difference, or no relationship between variables. In statistical testing, you start by assuming the null hypothesis is true and then use data to see whether there is enough evidence to reject it.

#### Alternative Hypothesis $H_1$ or $H_a$

An alternative hypothesis ( $H_1$  or  $H_a$ ) is the statement that says there is an effect, difference, or relationship. The alternative hypothesis is what you, the researcher, are actually trying to prove. It claims that a meaningful change, effect, or relationship is occurring in the data.

#### Example 1

A pharmaceutical company develops a new medication to treat high blood pressure. Standard treatment keeps the mean systolic blood pressure at 120 mmHg. Researchers want to test if the new drug changes the average blood pressure in *either* direction, lower or higher.

$H_0$ : The drug has no effect on average blood pressure and keeps 120 mmHg.

$H_1$ : The drug significantly alters average blood pressure.

#### Example 2

A school district introduces an after-school tutoring program. Historically, 75% of enrolled students graduate on time. The district wants to verify if the new program increases the graduation rate.

$H_0$ : The graduation rate remains 75% or lower.

$H_1$ : The graduation rate increases above 75%.

### Example 3

A factory packages cereal boxes labeled as containing 16 oz. Regulators audit the facility to test whether consumers are getting shortchanged (receiving less than 16 on average).

$H_0$ : The average net weight is at least 16 oz.

$H_1$ : The average net weight is strictly less than 16 oz.

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.

In hypothesis testing, a **parameter** is the specific numerical characteristic of an entire population that you want to investigate. Some of the testing parameters are given below. The testing parameter always has a specific value, called the claimed value.

#### 9.1.1: Parameter in Hypothesis Testing

|                                      | Parameter Symbol           | Example Claim                                                                         |
|--------------------------------------|----------------------------|---------------------------------------------------------------------------------------|
| <b>Population Mean</b>               | $\mu$ or $\mu_0$ ( $\mu$ ) | "The average weight of a smartphone is 180 grams."<br><br>$\mu = \mu_0 = 180$         |
| <b>Population Proportion</b>         | $p$ or $p_0$               | "More than 60% of voters support the new policy."<br><br>$p = p_0 \geq 0.60$          |
| <b>Population Standard Deviation</b> | $\sigma$ or $\sigma_0$     | "The variation in a manufacturing process 0.05 mm."<br><br>$\sigma = \sigma_0 = 0.05$ |

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 9.1.2: Mathematical Symbols Used in  $H_0$  and  $H_a$ :

|                         | $H_0$                                                                                                                 | $H_a$                                                                                                                |
|-------------------------|-----------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Sign contained          | =, $\leq$ , or $\geq$                                                                                                 | not equal ( $\neq$ ) <b>or</b> greater than ( $>$ ) <b>or</b> less than ( $<$ )                                      |
| Symbols                 | $H_0$ : parameter = claimed value<br>$H_0$ : parameter $\geq$ claimed value<br>$H_0$ : parameter $\leq$ claimed value | $H_a$ : parameter $\neq$ claimed value<br>$H_a$ : parameter $<$ claimed value<br>$H_a$ : parameter $>$ claimed value |
| What do we do on those? | Statement that is to be tested.                                                                                       | Statement which we are trying to find evidence to support                                                            |

$H_0$  always has a symbol with an equal in it and could be  $\leq$  or  $\geq$ .  $H_a$  never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. This practice is acceptable because we only decide whether to reject or not reject the null hypothesis.

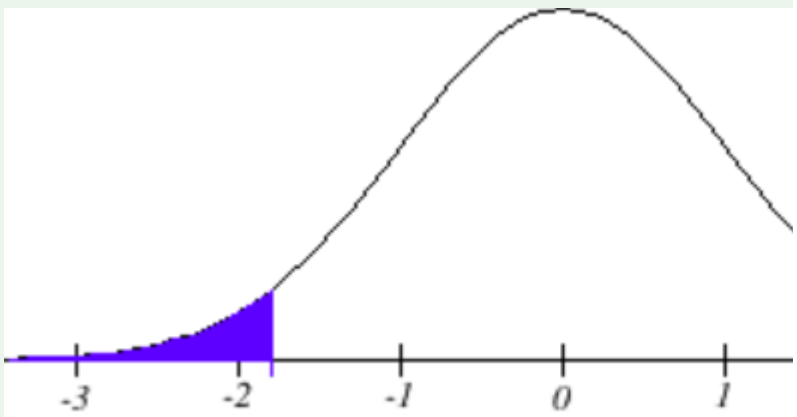
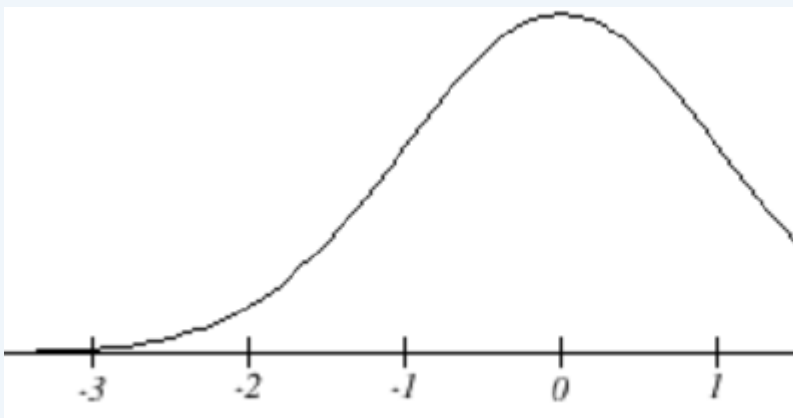
## Right-Tailed, Left-Tailed, and Two-Tailed Test

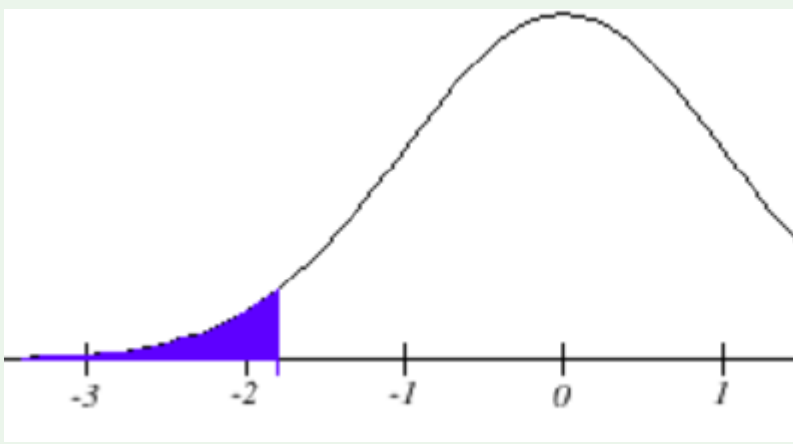
In hypothesis testing, the type of test is determined by the alternative hypothesis.

1. A right-tailed test is used when the alternative hypothesis claims that the population parameter is greater than ( $>$ ) a specified value or claimed value. The rejection region is located in the right tail of the distribution.
2. A left-tailed test is used when the alternative hypothesis claims that the population parameter is less than ( $<$ ) a specified value or claimed value. The rejection region is located in the left tail of the distribution.
3. A two-tailed test is used when the alternative hypothesis claims that the population parameter is different from (*neg*) a specified value. The rejection regions are split between both tails of the distribution.

### Comparisons of RTT, LTT, and TTT

#### RTT, LTT, and TTT

| Test Type                      | $H_a$ Sign | Rejection Region | Goal                                             | Graph                                                                                |
|--------------------------------|------------|------------------|--------------------------------------------------|--------------------------------------------------------------------------------------|
| <b>Left-Tailed Test (LTT)</b>  | $<$        | Left tail        | To see if the claimed value has <b>decreased</b> |   |
| <b>Right-Tailed Test (RTT)</b> | $>$        | Right tail       | To see if the claimed value has <b>increased</b> |  |

| Test Type                    | $H_a$ Sign | Rejection Region | Goal                                           | Graph                                                                              |
|------------------------------|------------|------------------|------------------------------------------------|------------------------------------------------------------------------------------|
| <b>Two-Tailed Test (TTT)</b> | $\neq$     | Both tails       | To see if the claimed value has <b>changed</b> |  |

#### Example 9.1.1: Two Tailed Test

State the null and alternative hypotheses in both words and symbolic form. Then identify the appropriate type of hypothesis test to be used.

"A drug company claims that its new medicine can lower cholesterol by 25%. Researchers conduct a medical trial to test whether the medicine really reduces cholesterol by that amount."

*Answer available at the end of this section.*

#### Example 9.1.2: Left Tailed Test

State the null and alternative hypotheses in both words and symbolic form. Then identify the appropriate type of hypothesis test to be used.

A factory packages cereal boxes labeled as containing 16 oz. Regulators audit the facility to test whether consumers are getting shortchanged (receiving less than 16 on average).

*Answer available at the end of this section.*

#### Example 9.1.3: Right Tailed Test

State the null and alternative hypotheses in both words and symbolic form. Then identify the appropriate type of hypothesis test to be used.

A school district introduces an after-school tutoring program. Historically, 75% of enrolled students graduate on time. The district wants to verify if the new program increases the graduation rate.

*Answer available at the end of this section.*

### Your Turn 9.1.3: Identify the Type of Test

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Conclusion in Hypothesis Test

The wording of the summary statement varies depending on which hypothesis the researcher claims is true. We should always set up the claim in the alternative hypothesis, since most of the time we are collecting evidence to show that a change has occurred. **Do not use the phrase “accept  $H_0$ ” since this implies that  $H_0$  is true.** The lack of evidence is not evidence of anything.

### Conclusion on Hypothesis Testing

There were only two possible correct answers for the decision step.

- i. Reject  $H_0$ .
- ii. Fail to reject  $H_0$ .

When you reject the null hypothesis, we say there is sufficient evidence to support the claim of the alternative hypothesis  $H_a$ .

When you do not reject the null hypothesis, we say there is NOT sufficient evidence to support the claim of the alternative hypothesis  $H_a$ .

### Your Turn 9.1.4: State Conclusion

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



Caution! If we fail to reject the null, this does not mean there was no change; it just means we have no evidence that a change occurred. The absence of evidence is not evidence of absence. On the other hand, we need to be careful when we reject the null hypothesis; we have not proved that there is a change.

When we reject the null hypothesis, there is only evidence that a change has occurred. Our evidence could have been false and led to an incorrect decision. If we use the phrase, “accept  $H_0$ ,” this implies that  $H_0$  was true, but we just do not have evidence that it is false. Hence, you will be marked incorrect for your decision if you use accept  $H_0$ , use instead “fail to reject  $H_0$ ,” or “do not reject  $H_0$ .”

### P-value and Level of Significance

The level of significance, denoted by  $\alpha$  (alpha), is the probability of rejecting the null hypothesis when the null hypothesis is actually true.

The level of significance is the cutoff used to decide whether the evidence against the null hypothesis is strong enough to reject it.

The  $p$ -value measures how strong the evidence is against the null hypothesis.

Note:  $\alpha$  value is set before the hypothesis test starts. The  $p$ -value is calculated from the data using technology.

### How to Draw a Conclusion Using $p$ -value and Alpha ( $\alpha$ )?

**Decision Rule:** Compare the  $p$ -value to the significance level  $\alpha$ :

- If  $p\text{-value} < \alpha$ , Reject  $H_0$ .
- If  $p\text{-value} \geq \alpha$  Don not reject  $H_0$ .

Interpretation

Small  $p$ -value  $\rightarrow$  strong evidence against the null hypothesis.

Large  $p$ -value  $\rightarrow$  weak evidence against the null hypothesis.

### Your Turn 9.1.5: Compare $p$ value and $\alpha$ to Draw Conclusion

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Test Statistic and Critical Values in Hypothesis Testing

The **critical value** (CV) is the boundary separating the **rejection region** from the **non-rejection region** in a hypothesis test.

CV notation

1. For  $z$ -test:  $\pm z_{\alpha}$  (For LLT and RTT),  $\pm z_{\frac{\alpha}{2}}$  (or TTT)
2. For  $t$ -test:  $\pm t_{\alpha}$  (For RTT and LTT),  $\pm t_{\frac{\alpha}{2}}$  (For TTT)

The **test statistic** is a number calculated from the sample data that is used to evaluate the null hypothesis. Test statistic is denoted by  $t_0 = t$  or  $z_0 = z$  depends on  $t$  test or  $z$  test.

**Decision Rule:** If the test statistic falls into the rejection region (beyond the critical value), reject the null hypothesis. See the following figures.

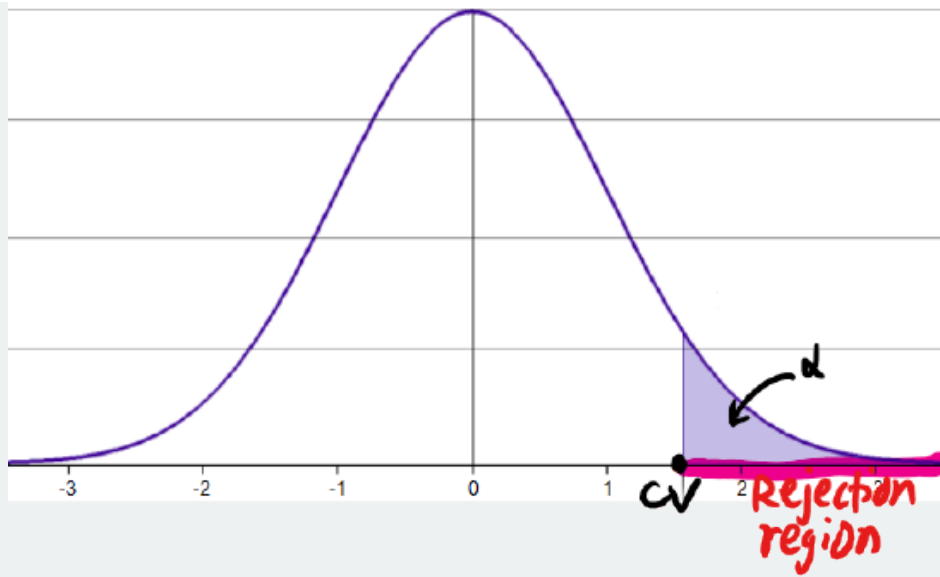


Figure 9.1.1: One Critical Value (RTT)

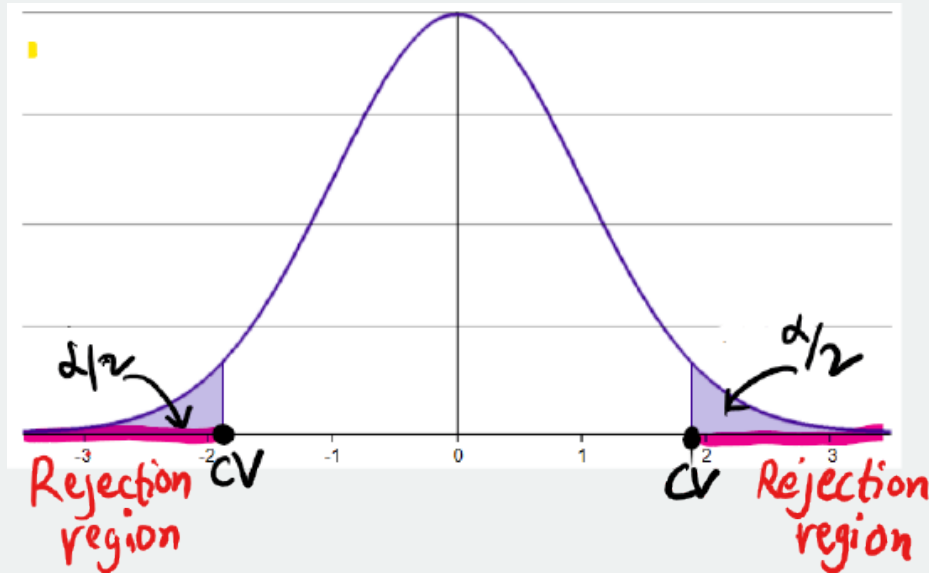


Figure 9.1.2: Two Critical Values (TTT)

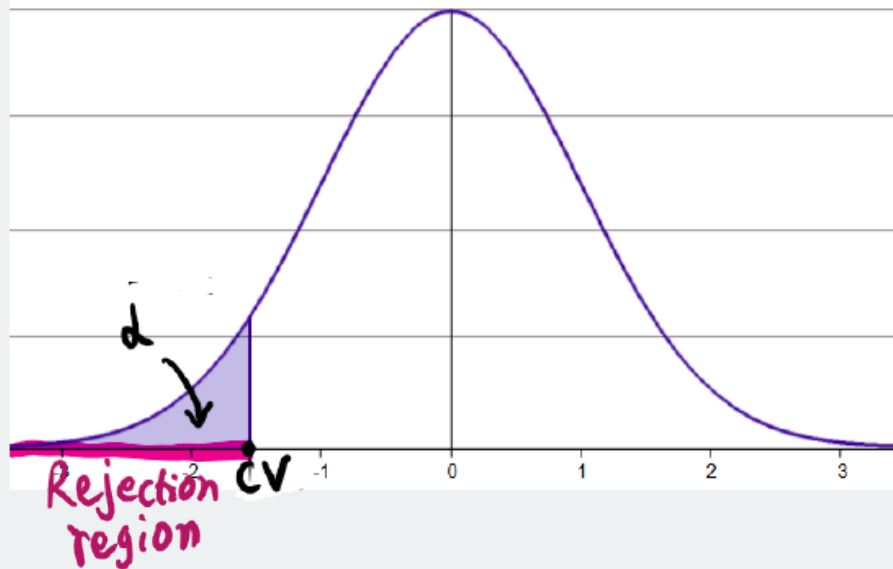


Figure 9.1.3: One Critical Value (LTT)

### Your Turn 9.1.6: Compare Test Statistic and CV to Draw Conclusion

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Errors in Hypothesis Testing

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:

### Error in Hypothesis Testing

| ACTION              | $H_0$ is True      | $H_1$ is True      |
|---------------------|--------------------|--------------------|
| Do not reject $H_0$ | Correct conclusion | Type II error      |
| Reject $H_0$        | Type I Error       | Correct conclusion |

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 error occurs with a particular probability. The Greek letters  $\alpha$  and  $\beta$  represent the probabilities.

- $\alpha$  = probability of a Type I error =  $P(\text{Type I error})$  = Probability of rejecting the null hypothesis when the null hypothesis is true.
- $\beta$  = probability of a Type II error =  $P(\text{Type II error})$  = Probability of not rejecting the null hypothesis when the null hypothesis is false.

$\alpha$  and  $\beta$  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 as close to 1 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 of Type I and Type II Errors

Suppose the null hypothesis: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital. Identify Type I and Type II errors.

$$H_0 : \text{Victim} = \text{alive}$$

$$H_a : \text{Victim} \neq \text{alive}$$

**Type I error:** Reject the victim is alive (the emergency crew thinks that the victim is dead) when, in fact, the victim is alive.

**Type II error:** Do not reject the victim is alive (the emergency crew does not know if the victim is alive) when, in fact, the victim is dead.

$\alpha = \text{Probability}$  that the emergency crew thinks the victim is dead when, in fact, he is really alive  $= P(\text{Type I error})$ .

$\beta = \text{Probability}$  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 more serious error is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

### Example 9.1.7: Type I and II Errors

The mean price of mid-sized cars in a region is \$57,900. A test is conducted to see if the claim is true.

*Answer available at the end of this section.*

### Your Turn 9.1.7: Identify Errors in Hypothesis Testing

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Attribution

MYOPEN MATH [www.myopenmath.com]

## Answers

### Example 9.1.1: Two Tailed Test

Here, the testing parameter is proportion  $p$  and the claimed value is 0.25.

$H_0$ : The drug reduces cholesterol by 25%.

$H_a$ : The drug does not reduce cholesterol by 25%.

The hypothesis in symbolic form is

$$H_0 : p = 0.25$$

$$H_a : p \neq 0.25$$

Because the alternative hypothesis uses the  $\neq$  symbol, this is a two-tailed test (TTT).

NOTE: In this example, the null hypothesis can also be written as  $p \geq 0.25$  because the company's claim will still be true if the medication lowers cholesterol more than 25%. The company will be accused of dishonesty if the drugs do not reduce cholesterol by 25%.

#### Example 9.1.2: Left Tailed Test

Here, the testing parameter is mean  $\mu$ , and the claimed value is 7 hours.

$H_0$ : The average net weight is at least 16 oz.

$H_1$ : The average net weight is strictly less than 16 oz.

The hypothesis in symbolic form is

$$H_0 : \mu = 16$$

$$H_a : \mu < 16$$

Because of the  $<$  symbol in the alternative hypothesis, this is a right-tailed test (LTT).

NOTE: In this example, the null hypothesis can also be written as  $p \geq 0.25$  because the company's claim will still be true if the net average weight is more than 16 oz. The company will be accused of dishonesty if the average weight is less than 16 oz.

#### Example 9.1.3: Right Tailed Test

Here, the testing parameter is mean  $p$ , and the claimed value is 0.75.

$H_0$ : The graduation rate remains 75% or lower.

$H_1$ : The graduation rate increases above 75%.

The hypothesis in symbolic form is

$$H_0 : \mu \leq 0.75$$

$$H_a : \mu > 0.75$$

Because of the  $>$  symbol in the alternative hypothesis, this is a two-tailed test (RTT).

Note: Since the alternative hypothesis has  $>$  symbol, it could be  $\leq$  symbol in the null hypothesis.

#### Example 9.1.7: Type I and II Errors

$$H_0 : p = \$57,900$$

$$H_a : p \neq \$57,900$$

**Type I:** Reject  $p = \$57,900$ , when  $p = \$57,900$  is true. In words, we decide that the mean price of mid-sized cars **is not** \$57,900, even though the mean price of mid-sized cars actually **is** \$57,900.

**Type II:** Do not reject  $p = \$57,900$ , when  $p \neq \$57,900$  is true. In words: we decide that the mean price of mid-sized cars **is** \$57,900, when the mean price of mid-sized cars actually **is not** \$57,900.

## 9.2: Test for the Population Proportion

Hypothesis testing for a proportion is a statistical method used to determine whether a population proportion differs significantly from a claimed or known value. It involves setting up a null hypothesis, which typically states that the population proportion equals a specific value, and an alternative hypothesis that suggests a difference. Using sample data, a test statistic is calculated and compared to a critical value or used to find a p-value. This process helps assess whether the observed sample proportion provides sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.

When you read a question, it is essential that you correctly identify the parameter of interest.

The parameter determines which model to use. Make sure you can distinguish between a question about a population means and a question about a population proportion.

The normal distribution can be used for a z-test for a proportion because, under certain conditions, the sampling distribution of the sample proportion approximates a normal distribution. This is based on the Central Limit Theorem, which states that as the sample size increases, the distribution of the sample proportion approaches normality, even if the population distribution is not normal.

### Criteria to Test Hypothesis about a Population Proportion

What are the conditions that must be met to test a hypothesis for a population proportion?

1. The sample must be a simple random sample
2.  $np \geq 10$  and  $n(1 - p) \geq 10$
3.  $N \geq 20n$

### Test Statistic $z_0$ or $z$ Formula

The formula for the *test statistic* is:

$$z \text{ or } z_0 = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

where  $n$  is the sample size.

$$\hat{p} = \frac{x}{n}$$

Where  $\hat{p}$  is the sample proportion (sometimes already given as a percentage) and  $p$  or  $p_0$  is the hypothesized **population proportion** or **claimed value**.

Also, you can find the test statistic using a calculator. See the video at the bottom of the page. I also explained in each example on this page.

### Statistically Significant

**Statistical significance** means that the results observed in a study or experiment are unlikely to have occurred by random chance alone.

In hypothesis testing, a result is considered **statistically significant** when the **p-value is less than the significance level,  $\alpha$** .

Note that the t-distribution is not used with proportions. We will use a standard normal z distribution for testing a proportion since this test uses the normal approximation to the binomial distribution (never use the t-distribution).

The p-value will always be a positive number between 0 and 1. The most important step in any method you use is setting up your null and alternative hypotheses. The critical values and p-value can be found using a standard normal distribution in the same way that we did for the one-sample z-test.

### How to Find the $p$ -value?

Also, you can use the Online Statistics Calculator to find the p-value. You need to use the Z-test of proportions to find the p-value. In the following example, you will know how to find the p-value using an online statistics calculator.

### Representation of the Critical Values and Rejection Region

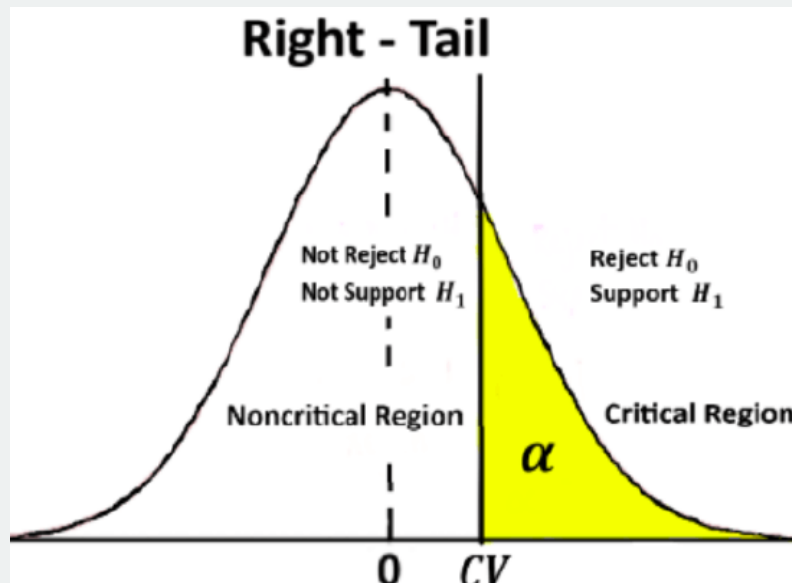


Figure 9.2.1: Right are on the tail is  $\alpha$

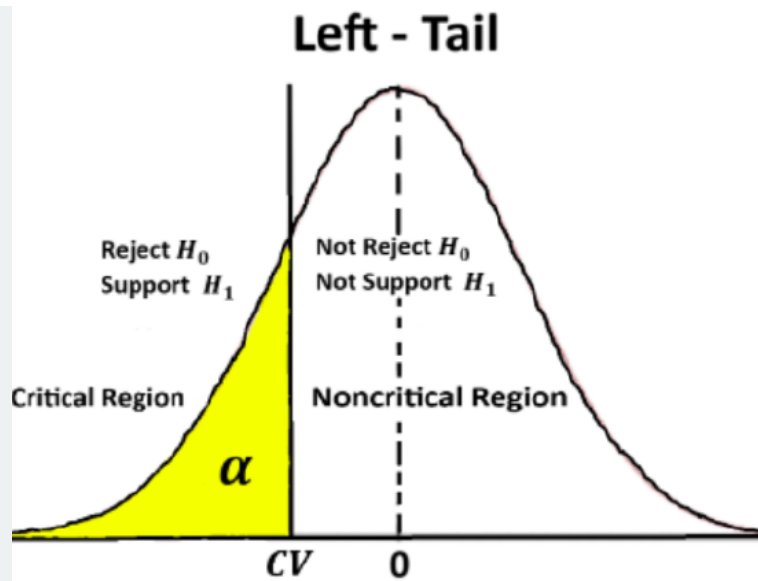


Figure 9.2.2: Left are on the tail is  $\alpha$

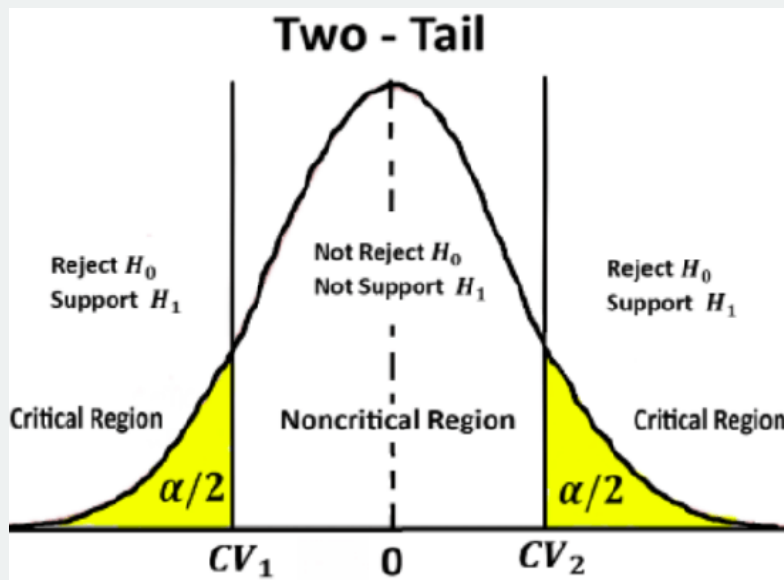


Figure 9.2.3: Total area on the tails is  $\alpha$

If you are doing a left-tailed z-test, the critical value (CV) will be negative. We also denote the left CV by  $-z_\alpha$ .

If you are performing a right-tailed z-test, the critical value (CV) will be positive. We also denote the right CV by  $z_\alpha$ .

If you were performing a two-tailed z-test, then your critical values would be both positive and negative.

We also denote both CV by  $\pm z_{\frac{\alpha}{2}}$ .

The attached video shows how to find the CV for RTT, LTT, and TTT.

## Your Turn 9.2.1: Find the critical value and Sketch Critical Region

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Two Approaches of Hypothesis Testing (p-Value and Classical)

There are two ways to test a hypothesis for a proportion. The following flowchart summarizes it.

### Testing Hypothesis for proportion (z-test)

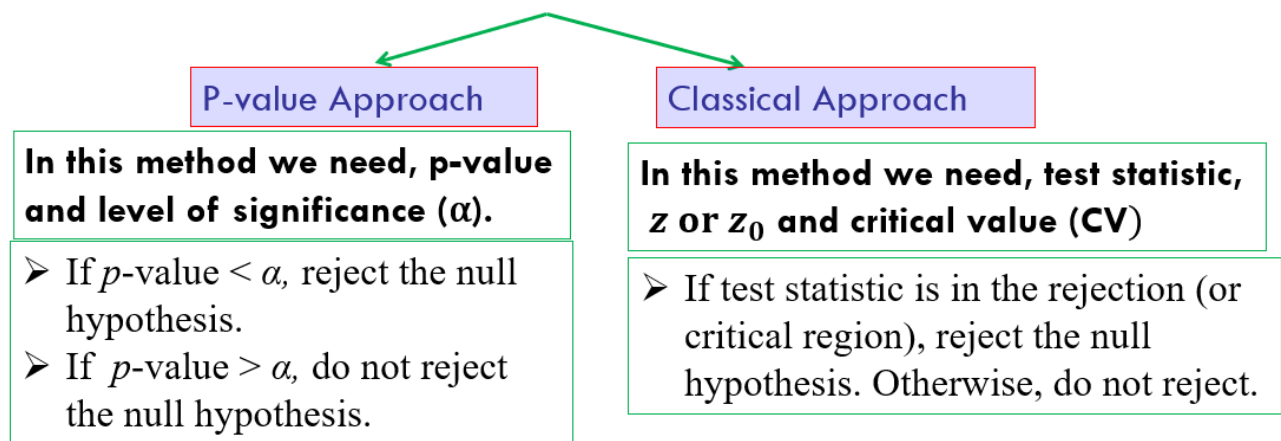


Figure 9.2.4: z-test for proportion p-value and classical approaches

## Steps to Test Hypothesis about Population Proportion

1. State the null and alternative hypotheses.
2. Verify that the requirements for a hypothesis test for a population proportion are satisfied.
3. Calculate the test statistic.
4. Find the p-value.
5. Calculate the critical value.
6. Draw and label the normal curve, showing the critical value and the test statistic.
7. State your decision (reject or fail to reject the null hypothesis).
8. Write a conclusion in the context of the problem.

Note: Steps 3, 4, and 5 can be found using an **Online Statistics Calculator**.

## Two-Tailed Test

In hypothesis testing, a **two-tailed test** is used to determine whether a sample statistic is significantly higher or lower than a hypothesized population parameter. Unlike a one-tailed test, which looks for a change in a single, specific direction, a two-tailed test splits the significance level ( $\alpha$ ) equally between both extremes of the probability distribution. For instance, if you are testing at a 5% significance level  $\alpha = 0.05$ , the rejection region is divided into two equal parts:

## Z Test Rejection Region

Two-tailed test, alpha = 0.050

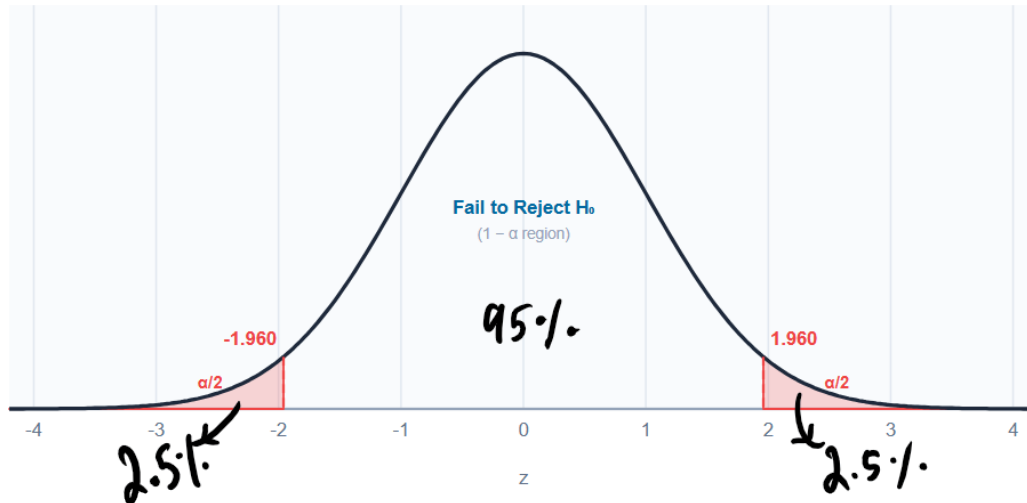


Figure 9.2.A: 95% of Total Area in the Middle of Two CV

2.5% of the total area in the upper tail and 2.5% of the total area in the lower tail. (See figure above)

This means the null hypothesis will be rejected if the sample evidence falls far enough into either end of the spectrum, indicating that the population proportion is simply *different* from what was hypothesized, rather than strictly greater than or less than it.

The following example shows how to test a two-tailed hypothesis using both classical and p-value methods.

### Example 9.2.2: Classical and P-value Method

It has been found that 85.6% of all enrolled college and university students in the United States are undergraduates. A random sample of 500 enrolled college students in a particular state revealed that 420 of them were undergraduates. Is there sufficient evidence to conclude that the proportion differs from the national percentage? Use  $\alpha = 0.05$ . Show that both methods of hypothesis testing yield the same results.

Answer available at the end of this section.

### Your Turn 9.2.2: Two Tailed Test

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



## Right-Tailed Test

A **right-tailed test** for a population proportion is a directional hypothesis test used to determine if the true population proportion,  $p$ , is significantly *greater* than a specified baseline or hypothesized value,  $p_0$ . In this scenario, the null hypothesis assumes the proportion is less than or equal to the baseline ( $H_0 : p \leq p_0$ ), while the alternative hypothesis claims it is strictly larger ( $H_1 : p > p_0$ ). Because the objective is to detect an increase, the entire critical region—or rejection zone—is located in the upper, right tail of the standard normal sampling distribution. If the observed sample proportion results in a test statistic ( $z$ -score) that falls far enough into this right tail, yielding a  $p$ -value less than the chosen significance level ( $\alpha$ ), the null hypothesis is rejected, providing

statistically significant evidence that the true population proportion has indeed increased.

### Example 9.2.3: Right Tailed Test

A Florida Community College claims that at least 30% of its students major in STEM fields. A student researcher believes that the proportion is greater than. To test this, the researcher surveys a random sample of 120 students and finds that 45 of them are majoring in STEM. Does the student researcher have enough evidence to support his claim with  $\alpha = 0.10$ ?

*Answer available at the end of this section.*

### Your Turn 9.2.3: Right Tailed Test

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Left -Tailed Test

A **left-tailed test** for a population proportion is a form of directional hypothesis testing used when you want to determine if the true population proportion,  $p$ , is significantly *less* than a specific claimed or baseline value,  $p_0$ . In this setup, the null hypothesis states that the proportion is greater than or equal to the baseline  $H_0 : p \geq p_0$ , while the alternative hypothesis strictly claims it is smaller  $H_1 : p < p_0$ .

Because the direction of interest points to the left, the entire critical region (the rejection zone) is located entirely in the lower, left tail of the sampling distribution. If the calculated sample proportion yields a test statistic ( $z$ -score) that falls deeply enough into this left tail—surpassing the critical value or resulting in a  $p$ -value less than the chosen significance level ( $\alpha$ )—the null hypothesis is rejected in favor of the alternative, indicating strong evidence that the true proportion has indeed decreased.

### Example 9.2.4: Left Tailed Test

A university claims that at least 80% of its students are satisfied with the quality of campus dining services. A student researcher believes that the actual proportion of satisfied students is less than 80%. The researcher randomly surveys 200 students and finds that 148 of them are satisfied with the campus dining services.

At the  $\alpha = 5\%$  significance level, test the claim that the population proportion of students who are satisfied with the campus dining services is less than 80%.

*Answer available at the end of this section.*

### Your Turn 9.2.4: Left Tailed Test

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



Please Practice This Question at least 3 Times

### Your Turn 9.2.5: Both $p$ -value and Classical Method

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this Interactive Content online](#)



Please practice following question: Your Turn 9.2.5 at least 3 times.

## Attributions

1. "8.4: Hypothesis Test for One Proportion" by Rachel Webb is licensed under [CC BY-SA 4.0](#)
2. "8.5: z-Test for a Proportion" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under [CC BY-SA 4.0](#)
3. <https://www.myopenmath.com/course/showlicense>

## Answers

### Example 9.2.2: Classical and P-value Method

At this point, you should be more comfortable with the steps of a hypothesis test and not have to number each step, but know what each step means.

**State the Hypothesis and Identify the Claim:** The keywords in this example, "proportion" and "differs," lead to the following hypotheses and claim.

$H_0 : p = 0.856$ ; Hypothesis to be tested

$H_1 : p \neq 0.856$ ; Hypothesis we are trying to find the evidence to support

Since the test is a two-tailed test (TTT) (because of  $\neq$ ), there are two critical values.

#### Check Conditions

The sample is assumed to be random, and

$$\begin{aligned} np &= 500 * 0.856 = 428 \geq 10 \\ n(1 - p) &= 50 * (1 - 0.856) = 72 \geq 10 \\ N &\geq 20n = 20 * 500 = 10,000 \end{aligned}$$

Note that in the last condition, the population size (the number of all US college students) is obviously greater than 10,000.

So all conditions are met

#### Use the online statistics Calculator to find the critical values

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 0$  and  $\sigma = 1$
3. Check two boxes  $\backslash(P(a$
4. Type 0.95 (middle area between two critical values) in the most right box at the bottom. The two numbers in the left two boxes are two critical values.

$$CV = z_{\frac{\alpha}{2}} = \pm 1.96$$

#### Compute the Test Statistic ( $z$ ) and P-value:

1. Put the cursor under Tests and click "z Test of a Proportion"
2. Type claimed value 0.856 in Null hypothesis box
3. Check  $\neq$  box
4. Type  $x = 420$  and  $n = 500$  and click calculate

At the bottom of a calculator window, you will find the p-value and the test statistic.

$$\begin{aligned} z &= z_0 = -1.02 \\ \text{p-value} &= 0.3081 \end{aligned}$$

We can also find the test statistic using the formula as shown below

$$\hat{p} = \frac{420}{500} = 0.84$$

Test statistic using formula

$$\begin{aligned}
 z &= \frac{\hat{p} - p}{\sqrt{\left(\frac{p(1-p)}{n}\right)}} \\
 &= \frac{0.84 - 0.856}{\sqrt{\left(\frac{0.856 \cdot (1 - 0.856)}{500}\right)}} \\
 &= -1.02
 \end{aligned}$$

**Make the Decision (Classical Method):** Draw and label the curve with the critical values and make the decision to reject or not reject the null hypothesis ( $H_0$ ). For  $\alpha = 0.05$ ,  $CV = z_{\frac{\alpha}{2}} = \pm 1.96$

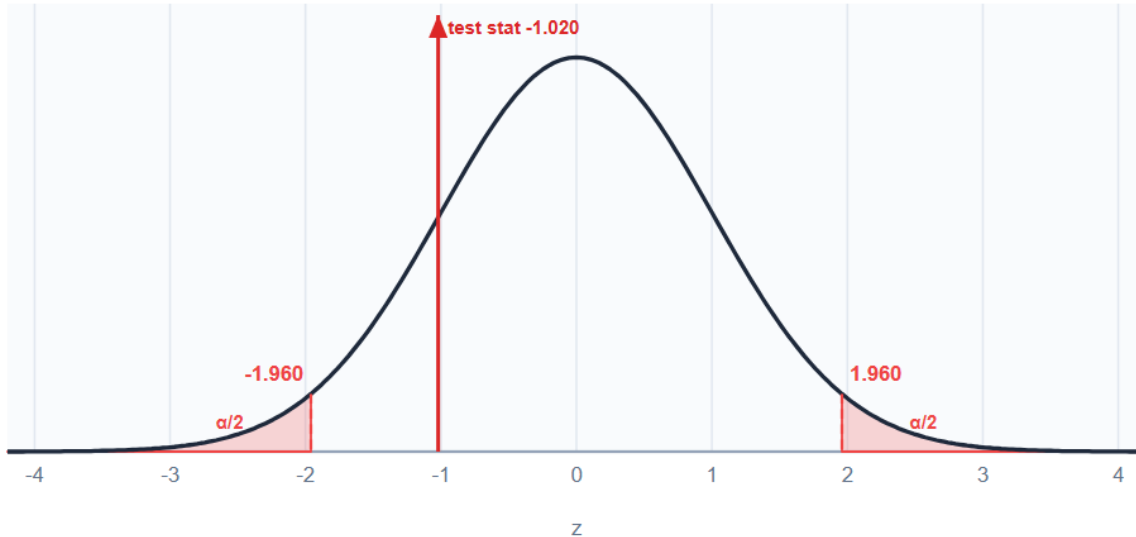


Figure 9.2.5: Rejection

Region with Critical Values  $\pm 1.96$  and Test Value of  $-1.02$ .

State the decision. Since the test statistic is not in the shaded rejection area, do not reject  $H_0$ . See Figure 9.2.5.

**Make the Decision (p-value Method):** Since the p-value  $> \alpha$ , the decision is not to reject  $H_0$ .

In this case, p-value = 0.3082  $> \alpha = 0.05$ : DO NOT reject  $H_0$ .

**Summarize the Results:** There is insufficient evidence to conclude that the proportion of undergraduates in college in this state differs from the national average of 85.6%.

#### Example 9.2.3: Right Tailed Test

##### State the Hypothesis and Identify the Claim:

The keywords in this example, "proportion" and "differs," lead to the following hypotheses and claim.

$H_0 : p = 0.30$ : Hypothesis to be tested

$H_1 : p > 0.30$ : Hypothesis we are trying to find the evidence to support

##### Check Conditions

$$\begin{aligned}
 np &= 120 * 0.30 = 36 \geq 10 \\
 n(1-p) &= 120 * (1 - 0.30) = 84 \geq 10 \\
 N &\geq 20n = 20 * 120 = 2400
 \end{aligned}$$

Note that in the last condition, the population size (the number of all Florida community college students) is obviously greater than 2400.

##### Find the Critical Value:

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 0$  and  $\sigma = 1$
3. Check two boxes  $P(x > a)$  and Calculate x-value
4. Type 0.10 (Right area from critical value) in the rightmost box at the bottom. The number in the middle box is the critical value.

$$CV = z_{\alpha} = 1.282$$

##### Compute the Test Statistic and p-value:

Before computing the test statistic, compute the sample proportion

1. Put the cursor under Tests and click "z Test of a Proportion"
2. Type claimed value 0.30 in Null hypothesis box
3. Check  $>$  box
4. Type  $x = 45$  and  $n = 120$  and click calculate

At the bottom of the calculator window, you will find the p-value and the test statistic.

$$z = z_0 = 1.793$$

$$p\text{-value} = 0.0365$$

You can also find test statistics using the formula  
The sample proportion is

$$\hat{p} = \frac{45}{120} = 0.375$$

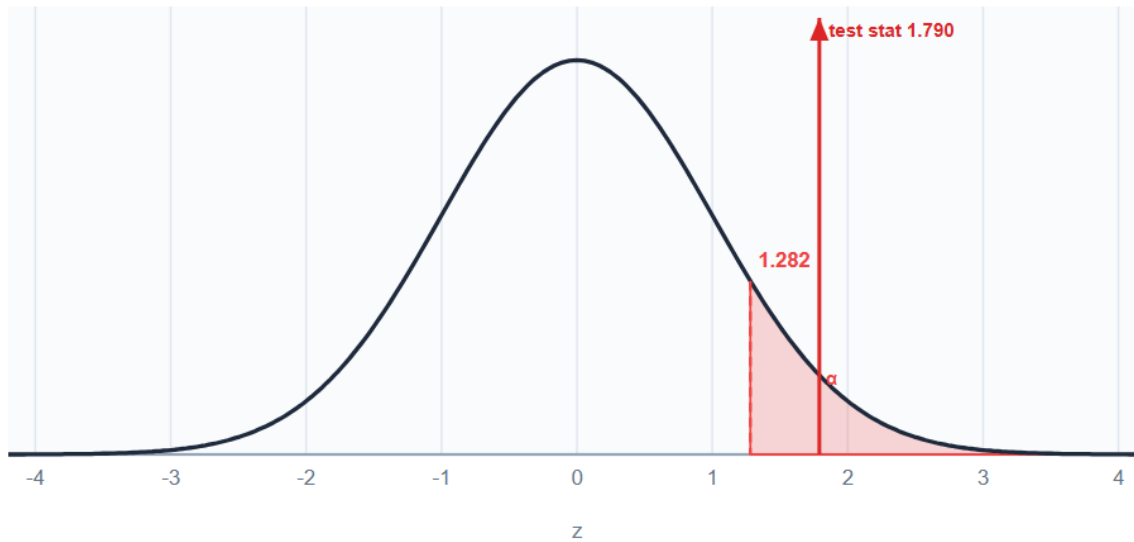
Test statistic using

$$\begin{aligned} z &= \frac{\hat{p} - p}{\sqrt{\left(\frac{p(1-p)}{n}\right)}} \\ &= \frac{0.375 - 0.3}{\sqrt{\left(\frac{0.3 \cdot 0.7}{120}\right)}} \\ &= 1.79 \end{aligned}$$

#### Make a decision (Classical Method):

For  $\alpha = 0.10$ ,  $CV = z_\alpha = 1.282$ .

Since the test statistic falls in the critical region, reject  $H_0$ . See figure 9.2.6.



Region with Critical Value 1.282 and Test Point 1.79.

Figure 9.2.6: Rejection

#### Make the Decision (p-value Method):

Since the p-value  $< \alpha$ , the decision is to reject  $H_0$ .

In this case, p-value = 0.0365  $< \alpha = 0.10$ : Reject  $H_0$

#### Draw the Conclusion:

There is enough evidence to conclude the proportion of STEM majors at the college is greater than 30%.

#### Example 9.2.4: Left Tailed Test

##### State the Hypothesis and Identify the Claim:

The keywords in this example, "less than," lead to the following hypotheses and claim.

$H_0 : p = 0.80$ : Hypothesis to be tested

$H_1 : p < 0.80$ : Hypothesis we are trying to find the evidence to support

##### Check Conditions

$$\begin{aligned} np &= 200 * 0.80 = 160 \geq 10 \\ n(1-p) &= 200 * (1 - 0.80) = 40 \geq 10 \\ N &\geq 20n = 20 * 200 = 4000 \end{aligned}$$

Note that in the last condition, we expect the population size (the number of all university students) to exceed 4000.

##### Find the Critical Value:

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 0$  and  $\sigma = 1$
3. Check two boxes  $\backslash P(x$
4. Type 0.05 (Left area from critical value) in the most right box at the bottom. The number in the middle box is the critical value.

$$CV = z_\alpha = -1.645$$

##### Compute the Test Statistic and P-value:

Before computing the test statistic, compute the sample proportion

1. Put the cursor under Tests and click "z Test of a Proportion"
2. Type claimed value 0.80 in Null hypothesis box
3. Check < box
4. Type  $x = 148$  and  $n = 200$  and click calculate

At the bottom of the calculator window, you will find the p-value and the test statistic.

$$z = z_0 = -2.121$$

$$p\text{-value} = 0.0169$$

The sample proportion is

$$\hat{p} = \frac{148}{200} = 0.74$$

Test statistic using the formula

$$\begin{aligned} z &= \frac{\hat{p} - p}{\sqrt{\left(\frac{p(1-p)}{n}\right)}} \\ &= \frac{0.74 - 0.80}{\sqrt{\left(\frac{0.80(1-0.80)}{200}\right)}} \\ &= -2.121 \end{aligned}$$

#### Make the Decision (Using Classical Method):

For  $\alpha = 0.05$ ,  $CV = z_\alpha = -1.645$

Since the test statistic falls in the rejection region, we reject  $H_0$ . See figure 9.2.7.

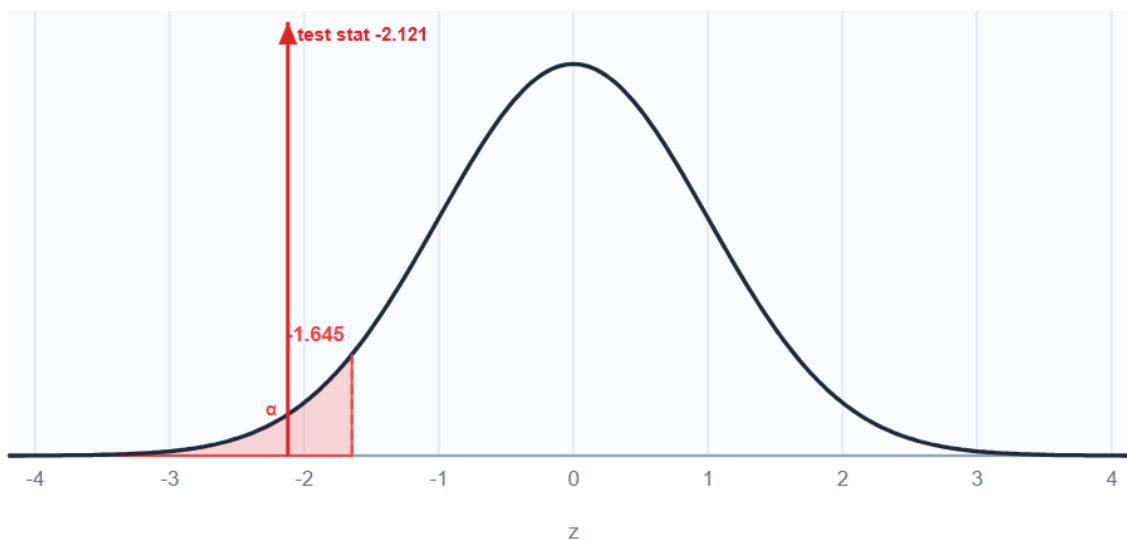


Figure 9.2.7: Rejection

Region with Critical Value  $-1.645$  and Test Point  $-2.12$ .

#### Make the Decision (Using p-value):

Since the  $p\text{-value} < \alpha$ , the decision is to reject  $H_0$ .

In this case,  $p\text{-value} = 0.0170 < \alpha = 0.05$  - reject  $H_0$ .

#### Draw the Conclusion:

There is sufficient evidence to support the claim that the population proportion of students who are satisfied with the campus dining services is less than 80%.

This page titled [9.2: Test for the Population Proportion](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [Back Matter](#) has no license indicated.

## 9.3: Test for the Population Mean

Hypothesis testing about a population mean is a formal inferential procedure used to evaluate whether sample data provides sufficient evidence to support a claim about an unknown population parameter  $\mu$ . Researchers establish a null hypothesis  $H_0$  asserting that the population mean equals a specific baseline value, paired with an alternative hypothesis  $H_1$  representing the competing claim. Depending on whether the population standard deviation is known or estimated from the sample, a  $z$ -statistic or  $t$ -statistic is calculated to measure how many standard errors the sample mean lies from the hypothesized value.

By comparing this test statistic—or its corresponding  $p$  value—against a predetermined significance level  $\alpha$ , researchers objectively decide whether to reject  $H_0$  or conclude that the sample evidence is insufficient to demonstrate a meaningful difference.

When you read a question text, it is essential that you correctly identify the parameter of interest.

The parameter determines which model to use. Make sure you can distinguish between a question about a population means and a question about a population proportion.

### Testing Hypothesis for Population Mean

The one-sample ( $t$ -test (or  $z$ -test) is a statistical method used to determine whether the average of a single sample is significantly different from a known or assumed population mean. It's especially useful when the population standard deviation is unknown, and the sample size is relatively small.

This test is commonly applied in a variety of fields. For example, in education, researchers might test whether the average test score of a new teaching method differs from the national average.

In healthcare, a doctor may want to check if a new treatment changes average blood pressure levels compared to a standard value. In manufacturing, a company might use a ( $t$ -test to determine whether the average weight of a product batch meets the target specification.

- **$z$ -Test:** Used when the population standard deviation ( $\sigma$ ) is **known**. It relies on the Standard Normal Distribution ( $z$ -distribution).
- **$t$ -Test:** Used when the population standard deviation ( $\sigma$ ) is **unknown**. It relies on Student's  $t$ -distribution and uses the sample standard deviation ( $s$ ) as an estimate.

The  $t$  and  $t$ -tests are easy to mix up. Sometimes, a standard deviation will be stated in the problem without specifying whether it is a population's standard deviation  $\sigma$  or the sample standard deviation  $s$ . If the standard deviation is in the same sentence that describes the sample or only raw data is given, then this would be  $s$ . When you only have sample data, use the  $t$ -test.

The table below highlights how **knowing** or **not knowing** the population standard deviation alters your statistical approach:

#### $t$ Test or $z$ Test?

$t$ -test or  $z$ -Test?

| Condition                | $\sigma$ is <b>known</b>                             | $\sigma$ is <b>unknown</b>                           |
|--------------------------|------------------------------------------------------|------------------------------------------------------|
| <b>Test Used</b>         | $Z$ -test of a mean                                  | $T$ - test of a mean                                 |
| <b>Critical Value</b>    | $z_\alpha, -z_\alpha,$ or $\pm z_{\frac{\alpha}{2}}$ | $t_\alpha, -t_\alpha,$ or $\pm t_{\frac{\alpha}{2}}$ |
| <b>Test Statistic</b>    | $z_0$ or $z$                                         | $t_0$ or $t$                                         |
| <b>Degree of Freedom</b> | Do not need here                                     | $df = n - 1$                                         |

## Criteria to Construct a Confidence Interval for a Population Mean

What are the conditions that you must look for to test a hypothesis for a population mean?

1. The sample must be a simple random sample
2. Is the population normal or not? If the population is not approximately normally distributed,  $n \geq 30$ .
3. Check to see if standard deviation is given or not to decide whether you use the  $z$ -test or the  $t$ -test.

## Test Statistic: $t$ or $z$ Formula

When  $\sigma$  is **unknown** (use  $t$  test for mean). In this case, the *test statistic* is

$$t \text{ or } t_0 = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

Where

Degree of freedom:  $df = n - 1$

Sample standard deviation:  $s$

Population mean:  $\mu$

Sample Mean:  $\bar{x}$

When  $\sigma$  is **known** (use  $z$  test for mean). In this case, the *test statistic* is

$$z \text{ or } z_0 = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

## Statistically Significant

**Statistical significance** means that the results observed in a study or experiment are unlikely to have occurred by random chance alone.

In hypothesis testing, a result is considered statistically significant when the p-value is less than the significance level, alpha  $\alpha$ .

## How to Find the $p$ -value?

Also, you can use the Online Statistics Calculator to find the p-value. You need to use the **Z Test of a Mean (or T Test of a Mean)** to find the p-value. In the following examples, you will know how to find the p-value using an online statistics calculator. Or you can watch the attached video.

# How to Find the Critical Value (CV) for RTT, LTT, and TTT?

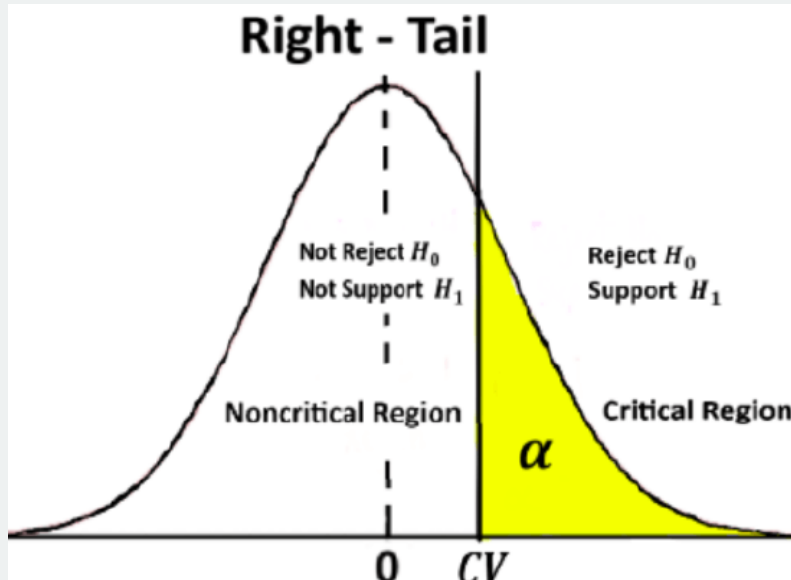


Figure 9.3.1: Right are on the tail is  $\alpha$

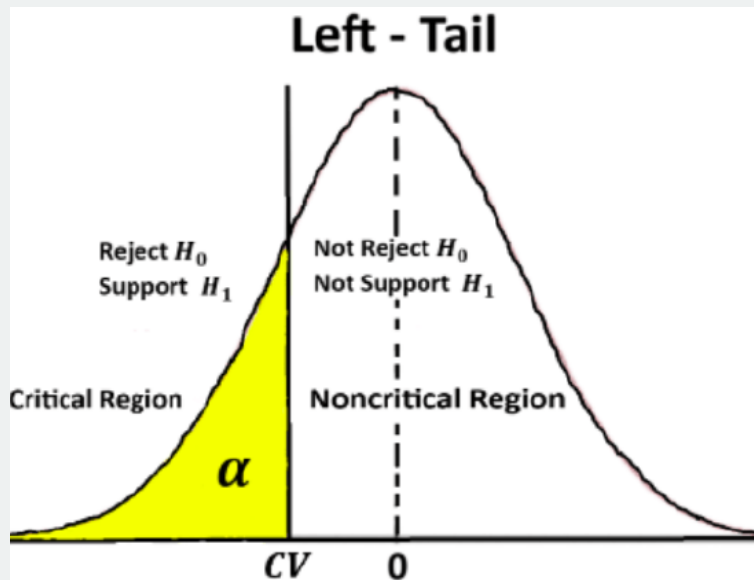


Figure 9.3.2: Left are on the tail is  $\alpha$

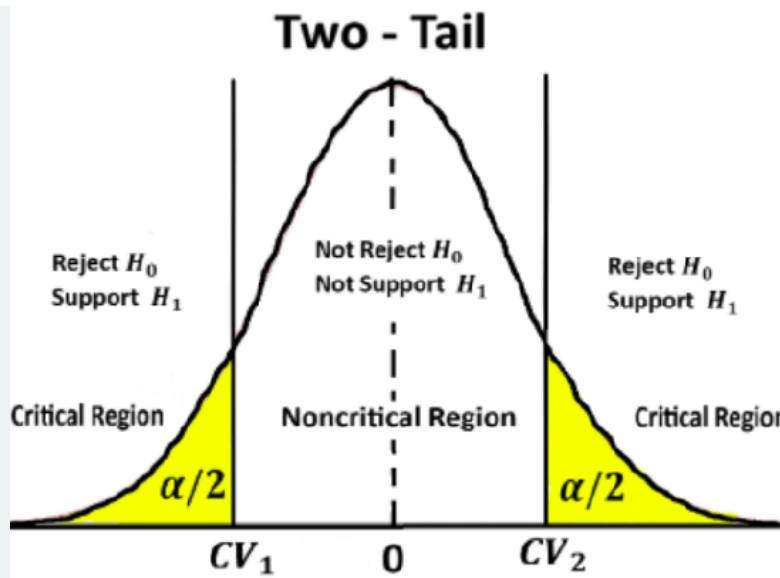


Figure 9.3.3: Total area on the tails is  $\alpha$

If you are doing a left-tailed z-test, the critical value (CV) will be negative. We also denote the left CV by  $-z_\alpha$  or  $-t_\alpha$ .

If you are performing a right-tailed z-test, the critical value (CV) will be positive. We also denote the right CV by  $z_\alpha$  or  $t_\alpha$ .

If you were performing a two-tailed z-test, then your critical values would be both positive and negative. We also denote both CV by  $\pm z_{\frac{\alpha}{2}}$  or  $\pm t_\alpha$ .

The attached video shows how to find the CV for RTT, LTT, and TTT.

### When do you reject the Null hypothesis?

There are two ways to test a hypothesis for a mean. The following flowchart summarizes it.

#### Testing Hypothesis for Mean (z-test or t-test)

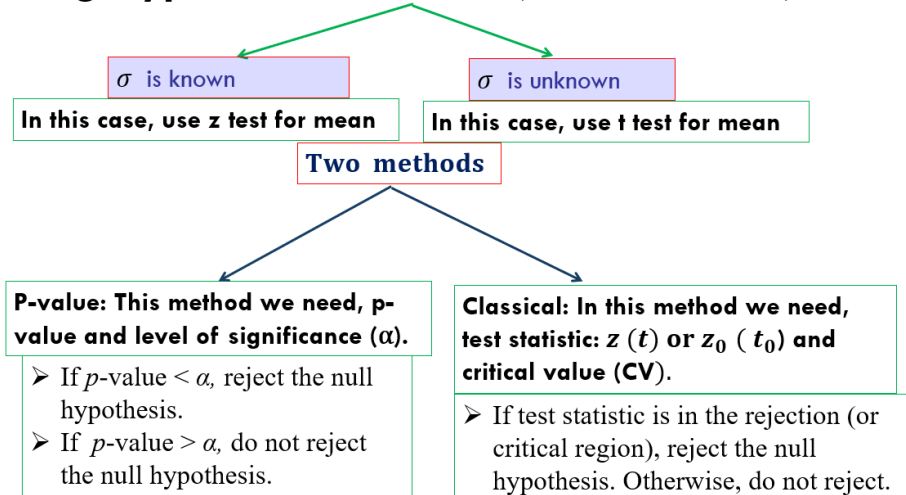


Figure 9.3.4: t-test or t-test p-value and classical Method

### Steps to Test Hypothesis about Population Parameter

- State the null and alternative hypotheses.
- Verify that the requirements for a hypothesis test for a population proportion are satisfied.
- Calculate the critical value(s).
- Calculate the test statistic.
- Find the p-value.
- Draw and label the normal curve, showing the critical value and the test statistic.
- State your decision (reject or fail to reject the null hypothesis).
- Write a conclusion in the context of the problem.

## Hypothesis Test for mean when $\sigma$ is Unknown ( $t$ Test of Mean)

As we know from figure 9.3.4, if  $\sigma$  is unknown, we use  $t$  test of mean.

### Example 9.3.1: TTT $\sigma$ is Unknown

The starting income of the recent STEM graduate is normally distributed. A career research analyst claims that recent graduates in STEM (Science, Technology, Engineering, and Mathematics) fields earn an average starting income of \$68,000. However, a university career center believes that this figure might be off and decides to conduct an investigation. They randomly surveyed a sample of 25 recent STEM graduates from their university. The results show that the average income of the sample is \$70,200, with a sample standard deviation of \$6,000. At  $\alpha = 0.10$ , test the career center's belief that the average starting income differs from \$68,000.

*Answer available at the end of this section.*

### Your Turn 9.3.1: Both $p$ -value and Classical Method

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 9.3.2: RTT $\sigma$ Unknown

With California's housing market known for its sky-high prices, a housing advocacy group claims that the average monthly rent statewide is \ (\$2,300\). However, a skeptical tenant rights organization believes that rent may be climbing even higher than reported. To test this belief, the organization surveyed a random sample of 36 California renters. The survey results show that the average monthly rent paid by the sample is \$2,480, with a sample standard deviation of \$500. At  $\alpha = 0.01$ , try to find the evidence to support the organization's claim that the average rent is greater than \$2,300.

*Answer available at the end of this section.*

### Your Turn 9.3.2: Both $p$ -value and Classical Method

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Hypothesis Test for Mean When $\sigma$ is Known ( $z$ Test of Mean)

As we know from figure 9.3.4, if  $\sigma$  is known, we use  $z$  test for mean.

### Example 9.3.3: LTT $\sigma$ is Known

The manufacturer of a particular brand of tires claims they average at least 81,000 km before needing to be replaced. From past studies of this tire, it is known that the population standard deviation is 13,100 km. A survey of tire owners was conducted to determine if the mileage of the tires is lower than 81,000 km. From the 35 tires sampled, the mean lifespan was 77,000 km. At  $\alpha = 10\%$ , can we prove that the data is inconsistent with the manufacturer's claim?

*Answer available at the end of this section.*

### Your Turn 9.3.3: Both $p$ -value and Classical Method

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 9.3.4: $t$ Test Data is Given

A real estate researcher is analyzing data on frequent movers. The researcher believes that people are becoming less mobile and settling down earlier in life. She claims that the average person used to live in is now less than 6 different homes by the age of 30. But she suspects that number is dropping.

To test this, she randomly surveyed 15 people under the age of 30 and recorded the number of different homes each person had lived in:

Data (number of homes):

4, 5, 3, 6, 5, 4, 5, 6, 3, 4, 4, 5, 4, 3, 4

Test the researcher's claim using  $\alpha = 0.05$ .

*Answer available at the end of this section.*

## Hypothesis Test for Mean When Data is Given

### Your Turn 9.3.4: $t$ -Test Data is Given

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



### Example 9.3.5: Understand the Problem

The label on a particular brand of cream of mushroom soup states that (on average) there is 870 mg of sodium per serving. A nutritionist would like to test if the average is more than the stated value. To test this, 13 servings of this soup were randomly selected, and the amount of sodium was measured. The sample mean was found to be 882.4 mg, and the sample standard deviation was 24.3 mg. Assume that the amount of sodium per serving is normally distributed. Test this claim using the traditional method of hypothesis testing. Use the  $\alpha = 0.05$  level of significance.

Read the text of this problem, and answer the following.

1. What are the null hypothesis and alternative hypothesis in this problem? State each one in words.
2. What type of test is this: a right-tailed test (RTT), left-tailed test (LTT), or two-tailed test (TTT)? Explain why.
3. What statistical test should you use: a  $t$ -test or a  $z$ -test? Is the population standard deviation known or unknown?
4. What do the numerical values given in the problem represent?
5. How many critical values are there for this test?

*Answer available at the end of this section.*

Please practice the following question at least 3 times.

### Your Turn 9.3.5: Both $p$ -value and Classical Method

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

[View this Interactive Content online](#)



## Attributions

1. "8.3: Hypothesis Test for One Mean"(opens in new window) by Rachel Webb(opens in new window) is licensed under [CC BY-SA 4.0](#)(opens in new window)
2. "9.E: Hypothesis Testing with One Sample (Exercises)"(opens in new window) by OpenStax(opens in new window) is licensed under [CC BY 4.0](#)(opens in new window)
3. This page is modified from the page by Toros Berberyan, Tracy Nguyen, and Alfie Swan, which is licensed under [CC BY-SA 4.0](#) (Thank you to the author).
4. <https://www.myopenmath.com/course/sh...836524-1834176>

## Answers

### Example 9.3.1: TTT $\sigma$ is Unknown

#### State the Hypotheses and Identify the Claim:

The statement "the income differs from \$68,000" must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

$H_0 : \mu = 68,000$ : Hypothesis to be tested

$H_1 : \mu \neq 68,000$ : (Hypothesis we are trying to find the evidence to support)

This is a two-tailed test.

#### Check Conditions

The sample is assumed to be random, and the population standard deviation  $\sigma$  is unknown. We use the  $t$  test here. The population is normal

#### Find the Critical Values: Use Online Statistics Calculator

1. Put the cursor under Distribution and click "Student's T"
  2. Type  $\bar{x} = 0$ ,  $s = 1$ , and  $df = 25 - 1 = 24$
  3. Check two boxes  $\backslash P(a$
  4. Type 0.90 (middle area between two critical values, because  $(CL = 1 - \alpha = 1 - 0.10 = 0.90)$  in the most right box at the bottom.
- The two numbers in the left two boxes are two critical values.

$$CV = t_{\frac{\alpha}{2}} = \pm 1.711$$

#### Compute the Test Statistic ( $t$ ) and p-value:

1. Put the cursor under Tests and click "T Test of a Mean"
  2. Type claimed value 68000 in Null hypothesis box
  3. Check  $\neq$  box and Input from Stat box.
  4. Type  $\bar{x} = 70,200$ ,  $s = 6,000$ , and  $n = 25$  and click calculate
- At the bottom of the calculator window, you will find the p-value and the test statistic.

$$t = t_0 = 1.833$$

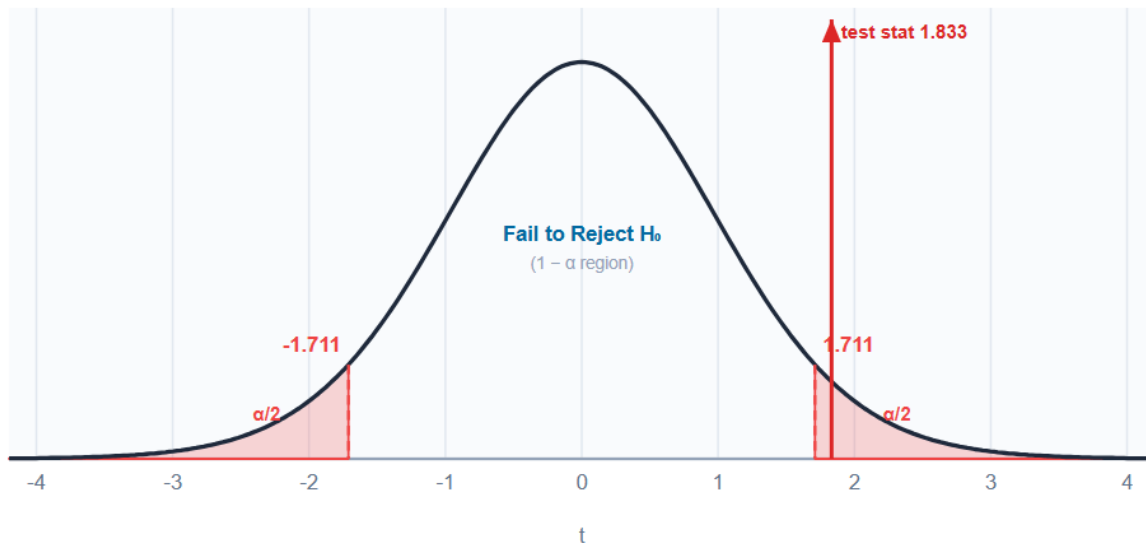
$$p\text{-value} = 0.0792$$

We can also find the test statistic using the formula as shown below

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{70,200 - 68,000}{\frac{6000}{\sqrt{25}}} \\ &= 1.833 \end{aligned}$$

#### Make the Decision (Classical Method):

Since the test statistic  $t = 1.833$  is in the critical region, we should reject  $H_0$ .



Values  $\pm 1.711$  and Test Value of 1.833.

Figure 9.3.5: with Critical

#### Make the Decision (p-value Method):

Since the  $p\text{-value} < \alpha$ , the decision is to reject  $H_0$

In this case,  $p\text{-value} = 0.0792 < \alpha = 0.10$ : Reject  $H_0$ .

### Draw the Conclusion:

There is enough evidence to prove the University Career Center's claim that the starting income of the recent STEM graduate might be off (not equal to \$68,000).

#### Example 9.3.2: RTT $\sigma$ Unknown

**State the Hypotheses and Identify the Claim:** The statement "the rent is greater than \$2,300" must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

$H_0 : \mu = 2,300$ : Hypothesis to be tested

$H_1 : \mu > 2,300$ : Hypothesis we are trying to find the evidence to support

This is a right-tailed test.

#### Find the Critical Values: Use Online Statistics Calculator

1. Put the cursor under distribution and click "Student's T"
2. Type  $\bar{x} = 0$ ,  $s = 1$ , and  $df = 36 - 1 = 35$
3. Check two boxes  $P(x > a)$  and Calculate x-value
4. Type 0.01 (right area form the critical value) in the most right box at the bottom. The left number in the first box is a critical value.

$$CV = t_{\frac{\alpha}{2}} = 2.438$$

#### Compute the Test Statistic ( $t$ ) and p-value:

1. Put the cursor under Tests and click "T Test of a Mean"
2. Type claimed value 2300 in Null hypothesis box
3. Check  $>$  box and Input from Stat box.
4. Type  $\bar{x} = 2480$ ,  $s = 500$ , and  $n = 36$  and click calculate

At the bottom of the calculator window, you will find the p-value and the test statistic.

$$t = t_0 = 2.16$$

$$\text{p-value} = 0.0189$$

We can also find the test statistic using the formula as shown below

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{2480 - 2300}{\frac{500}{\sqrt{36}}} \\ &= 2.16 \end{aligned}$$

#### Make the Decision (Using Classical Method):

Since the test statistic does not fall in the rejection region, we do not reject  $H_0$ . See figure 9.3.6.

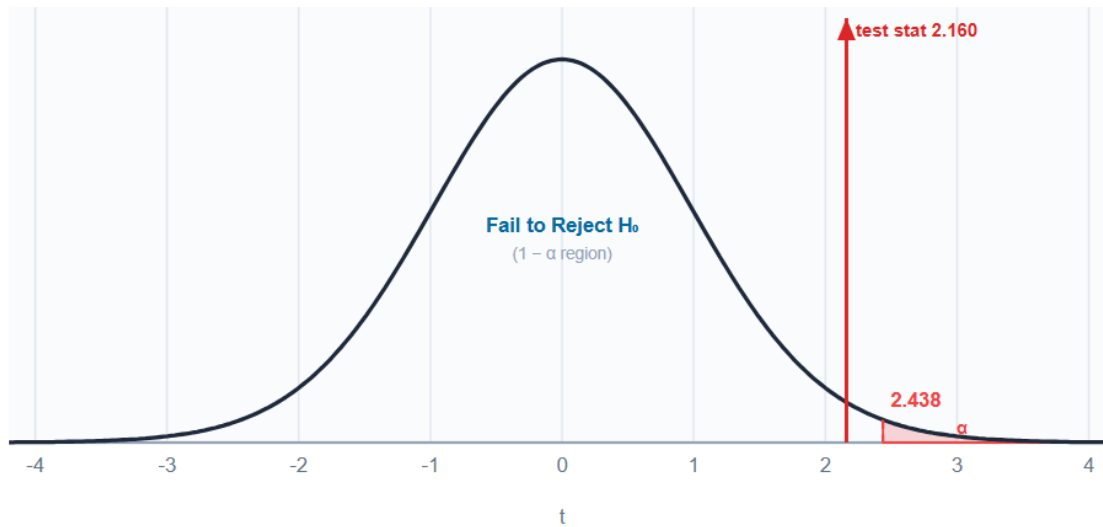


Figure 9.3.6: Critical Value

2.483 and Test Value of 2.16.

**Make the Decision (Using p-value):** Since the p-value  $> \alpha$ , the decision is to reject  $H_0$ .

In this case, p-value = 0.0189  $>$   $\alpha = 0.01$ : Do not reject  $H_0$ .

**Draw the Conclusion:** There is insufficient evidence to support the organization's claim that the average monthly rent in California exceeds \$2,300.

#### Example 9.3.3: LTT $\sigma$ is Known

##### State the Hypotheses and Identify the Claim:

The statement "the mileage of the tire is lower than 81,000" must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

$H_0 : \mu = 81,000$ : Hypothesis to be tested

$H_1 : \mu < 81,000$ : Hypothesis we are trying to find the evidence to support

This is a left-tailed test. Since  $\sigma$  is known, use ( $z$ ) test of a mean.

### Find the Critical Values: Use Online Statistics Calculator

1. Put the cursor under distribution and click "Normal"
2. Type  $\mu = 0$  and  $\sigma = 1$ .
3. Check two boxes  $\backslash P(x)$
4. Type 0.10 (left area from the critical value) in the rightmost box at the bottom. The number in the middle box is a critical value.

$$CV = z_{\frac{\alpha}{2}} = -1.282$$

### Compute the Test Statistic ( $z$ ) and p-value:

1. Put the cursor under Tests and click "Z Test of a Mean"
  2. Type claimed value 81,000 in Null hypothesis box
  3. Check  $\neq$  box and Input from Stat box.
  4. Type  $\bar{x} = 77,000$ ,  $\sigma = 13,100$ , and  $n = 35$  and click calculate
- At the bottom, you will find the p-value and the test statistic.

$$z = z_0 = -1.806$$

$$p\text{-value} = 0.0354$$

We can also find the test statistic using the formula as shown below

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{77,000 - 81,000}{\frac{13,100}{\sqrt{35}}} \\ &= -1.806 \end{aligned}$$

### Make the Decision (Using Classical Method):

For  $\alpha = 0.10$ ,  $CV = t_{\alpha} = -1.806$

Since the test statistic falls in the rejection region, we reject  $H_0(\mu = 81,000)$ . See figure 9.3.7.

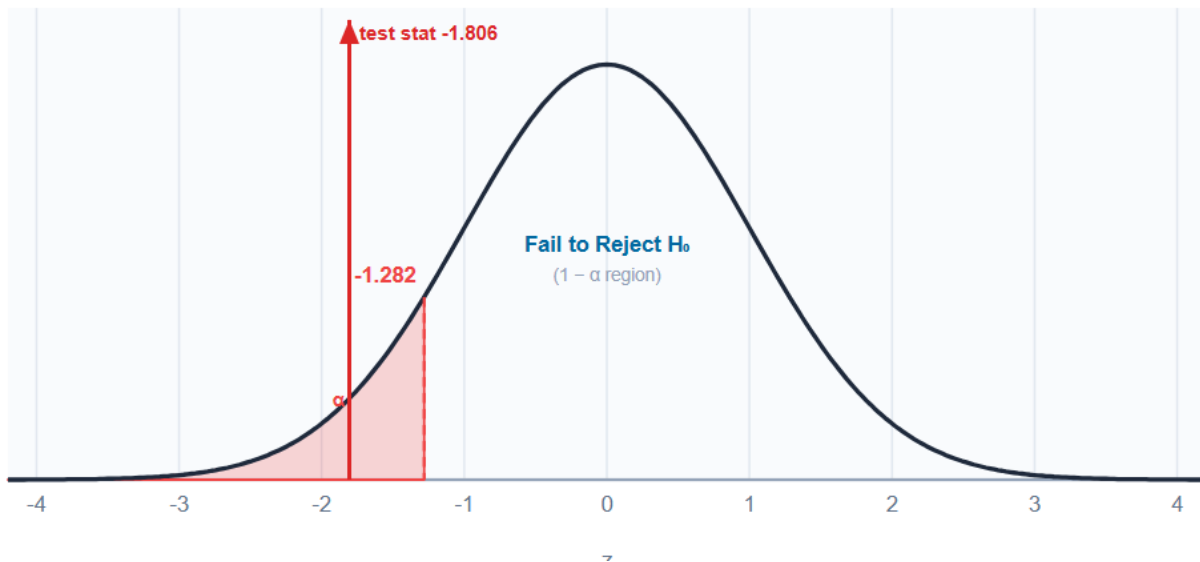


Figure 9.3.7: Critical Value  $-1.282$  and Test Value of  $-1.806$ .

### Make the Decision (Using p-value):

Since the p-value  $< \alpha$ , the decision is to reject  $H_0$ .

In this case, p-value  $= 0.0354 < \alpha = 0.10$ : Reject  $H_0(\mu = 81,000)$ .

### Draw the Conclusion:

There is sufficient evidence to conclude that the tires last fewer kilometers than the claimed 81,000 km.

### Example 9.3.4: $t$ Test Data is Given

#### State the hypotheses and identify the claim:

The statement "the average is less than 6" must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

$H_0 : \mu = 6$ : Hypothesis to be tested

$H_1 : \mu < 6$ : Hypothesis we are trying to find the evidence to support

This is a left-tailed test. Since  $\sigma$  is unknown, use  $t$  test of a mean

### Find the Critical Values: Use Online Statistics Calculator

1. Put the cursor under distribution and click "Student's T"
2. Type  $\bar{x} = 0$ ,  $s = 1$ , and  $df = 15 - 1 = 14$
3. Check two boxes  $\backslash P(x)$
4. Type 0.01 (left area from the critical value) in the rightmost box at the bottom. The left number in the first box is a critical value.

$$CV = t_{\alpha} = -1.761$$

**Compute the Test Statistic ( $z$ ) and p-value:**

1. Click Class Editor and put all the data in L1
2. Put the cursor under Tests and click "T Test of a Mean"
3. Type claimed value 6 in Null hypothesis box
4. Check  $<$  box and Input from Stat box.
5. Type L1 in the list box. (L1 should already be there) and press calculate

At the bottom of the calculator window, you will find the p-value and the test statistic.

$$t = t_0 = -6.614$$

$$p\text{-value} = 0.000$$

We can also find the test statistic using the formula as shown below

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{4.333 - 6}{\frac{0.9759}{\sqrt{15}}} \\ &= -6.614 \end{aligned}$$

**Make the Decision (Using Classical Method):**

For  $\alpha = 0.05$ ,  $CV = t_{\alpha} = -1.761$

Since the test statistic ( $-6.614$ ) falls in the rejection region, we reject  $H_0(\mu = 6)$ . See figure 9.3.8.



Figure 9.3.8: The t-value  $-6.14$  Falls in the Critical Region.

**Make the Decision (Using p-value):**

Since the p-value  $< \alpha$ , the decision is to reject  $H_0$ .

In this case, p-value  $= 0 < \alpha = 0.05$ ; Reject  $H_0(\mu = 6)$ .

**Draw the Conclusion:**

There is enough evidence to support the researcher's claim that the average number of homes people have lived in is now less than 6.

**Example 9.3.5: Understand the Problem**

1. Hypothesis in words

Null: Particular brand of cream of mushroom soup has 870 mg of sodium per serving.

Alternative: That brand of cream of mushroom soup has more than 870 mg of sodium per serving.

2. This is a right-tailed test with the claim in the alternative hypothesis. Because of **more than** one word in the alternative hypothesis.

3. Since  $\sigma$  is unknown, we use  $t$  test of mean.

- 4.

$\mu = 870$  : population mean or claimed value

$\bar{x} = 882.4$  : Sample mean

$s = 24.3$  : Sample Standard Deviation

$n = 13$  : Sample Size

$df = n - 1 = 13 - 1 = 12$  : Degree of Freedom

$\alpha = 0.05 = 5\%$  : Level of Significance

5. There is one positive critical value since it is a right-tailed test

---

This page titled [9.3: Test for the Population Mean](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

- [Back Matter](#) has no license indicated.
- [Front Matter](#) has no license indicated.

# INDEX

## A

**addition rule**  
     4: Probability  
**alternative hypothesis**  
     9: Hypothesis Testing  
     9.1: Language of Hypothesis Testing  
     9.2: Test for the Population Proportion  
     9.3: Test for the Population Mean  
**applications**  
     6: The Normal Distribution  
**area under curve**  
     6.2: Normal Distribution and Applications  
**at least one rule**  
     4.3: Independence and Multiplication Rules  
**average**  
     2.1: Measures of Center of the Data

## B

**bar graph**  
     1.2: Displaying Data  
**bell curve**  
     6: The Normal Distribution  
     6.1: Properties of the Normal Distribution  
**binomial**  
     5: Discrete Probability Distributions  
**binomial experiment**  
     5.2: Binomial Probability Distributions  
**binomial formula**  
     5.2: Binomial Probability Distributions  
**bivariate data**  
     3.1: Scatter Diagrams and Correlation  
**box plot**  
     2.3: Measures of Location of the Data

## C

**categorical data**  
     1.2: Displaying Data  
**categorical variables**  
     4.2: The Addition and Complement Rules  
**central limit theorem**  
     7: Sampling Distributions  
     7.1: The Sampling Distribution of the Sample Mean  
**central tendency**  
     2: Numerical Summary of Data- Center, Variation, and Location  
     2.1: Measures of Center of the Data  
**coin flip**  
     4.3: Independence and Multiplication Rules  
**conditional probability**  
     4.2: The Addition and Complement Rules  
**confidence interval**  
     8.1: Estimating Population Proportion  
     8.2: Estimating Population Mean  
     9.3: Test for the Population Mean  
**confidence intervals**  
     8: Estimating the Value of a Parameter  
     9: Hypothesis Testing  
**confidence level**  
     8.1: Estimating Population Proportion

**contingency table**  
     4.2: The Addition and Complement Rules  
**continuous**  
     5: Discrete Probability Distributions  
**continuous data**  
     5.1: Discrete Probability Distributions  
**correlation**  
     3: Linear Regression and Correlation  
     3.1: Scatter Diagrams and Correlation  
**correlation coefficient**  
     3.2: Least Squares Regression Line  
**critical evaluation**  
     1: Introduction to Data  
**critical thinking**  
     1.1: Data, Samples, and Populations  
**critical value**  
     8.1: Estimating Population Proportion  
     9.2: Test for the Population Proportion  
**cumulative frequency**  
     1.2: Displaying Data  
**cumulative probability**  
     6.2: Normal Distribution and Applications  
**D**  
**data**  
     1: Introduction to Data  
     5: Discrete Probability Distributions  
**data analysis**  
     1.1: Data, Samples, and Populations  
     2.1: Measures of Center of the Data  
     2.2: Measures of Spread of the Data  
     2.3: Measures of Location of the Data  
     5.1: Discrete Probability Distributions  
     6.2: Normal Distribution and Applications  
**data distribution**  
     2: Numerical Summary of Data- Center, Variation, and Location  
     2.1: Measures of Center of the Data  
     2.3: Measures of Location of the Data  
**data fit**  
     3: Linear Regression and Correlation  
**data interpretation**  
     1.1: Data, Samples, and Populations  
     2.3: Measures of Location of the Data  
**data visualization**  
     1.2: Displaying Data  
     3.1: Scatter Diagrams and Correlation  
**decision making**  
     9.1: Language of Hypothesis Testing  
**density function**  
     8: Estimating the Value of a Parameter  
**dependent events**  
     4.3: Independence and Multiplication Rules  
**dependent variable**  
     3: Linear Regression and Correlation  
     3.2: Least Squares Regression Line  
**descriptive statistics**  
     2: Numerical Summary of Data- Center, Variation, and Location  
     2.2: Measures of Spread of the Data  
**die roll**  
     4.3: Independence and Multiplication Rules

**discrete**  
     5: Discrete Probability Distributions  
**discrete data**  
     5.1: Discrete Probability Distributions  
**discrete probability distribution**  
     5.2: Binomial Probability Distributions  
**distributions**  
     5: Discrete Probability Distributions

## E

**empirical probabilities**  
     4: Probability  
**empirical rule**  
     2.2: Measures of Spread of the Data  
**event**  
     4.1: Fundamental Concepts of Probability  
**events**  
     4: Probability  
**evidence**  
     9: Hypothesis Testing  
**expected value**  
     5.1: Discrete Probability Distributions  
**experiment**  
     4.1: Fundamental Concepts of Probability  
**experimental**  
     5: Discrete Probability Distributions  
**experimental procedure**  
     4.1: Fundamental Concepts of Probability  
**experiments**  
     4: Probability  
**explanatory variable**  
     3.1: Scatter Diagrams and Correlation

## F

**failure**  
     5.2: Binomial Probability Distributions  
**fixed number of trials**  
     5.2: Binomial Probability Distributions  
**frequency**  
     4.2: The Addition and Complement Rules  
**frequency table**  
     1.2: Displaying Data

## G

**gaussian distribution**  
     6.1: Properties of the Normal Distribution

## H

**histogram**  
     1.2: Displaying Data  
**hypotheses**  
     9.1: Language of Hypothesis Testing  
**hypothesis testing**  
     9: Hypothesis Testing  
     9.1: Language of Hypothesis Testing  
     9.2: Test for the Population Proportion  
     9.3: Test for the Population Mean

## I

### independent events

4.3: Independence and Multiplication Rules

### independent trials

5.2: Binomial Probability Distributions

### independent variable

3: Linear Regression and Correlation  
3.2: Least Squares Regression Line

### inference

7.2: The Sampling Distribution of Sample Proportion

### inferential statistics

1.1: Data, Samples, and Populations  
7.1: The Sampling Distribution of the Sample Mean

### interquartile range

2.3: Measures of Location of the Data

## L

### law of large numbers

7: Sampling Distributions

### least squares regression line

3: Linear Regression and Correlation  
3.2: Least Squares Regression Line

### level of significance

9.1: Language of Hypothesis Testing

### linear correlation coefficient

3: Linear Regression and Correlation  
3.1: Scatter Diagrams and Correlation

### linear regression

3.2: Least Squares Regression Line

## M

### margin of error

8.1: Estimating Population Proportion  
8.2: Estimating Population Mean

### mean

2: Numerical Summary of Data- Center, Variation, and Location  
2.1: Measures of Center of the Data  
2.2: Measures of Spread of the Data  
5.1: Discrete Probability Distributions  
6: The Normal Distribution  
6.1: Properties of the Normal Distribution  
6.2: Normal Distribution and Applications  
7.2: The Sampling Distribution of Sample Proportion

### measures of center

2: Numerical Summary of Data- Center, Variation, and Location  
2.1: Measures of Center of the Data

### median

2: Numerical Summary of Data- Center, Variation, and Location  
2.1: Measures of Center of the Data  
2.3: Measures of Location of the Data

### mode

2: Numerical Summary of Data- Center, Variation, and Location  
2.1: Measures of Center of the Data

### multiplication rule

4.3: Independence and Multiplication Rules

### mutually exclusive events

4: Probability

## N

### normal distribution

2.2: Measures of Spread of the Data  
6: The Normal Distribution  
6.1: Properties of the Normal Distribution  
6.2: Normal Distribution and Applications  
7.1: The Sampling Distribution of the Sample Mean  
8: Estimating the Value of a Parameter

### null hypothesis

9: Hypothesis Testing  
9.1: Language of Hypothesis Testing  
9.2: Test for the Population Proportion  
9.3: Test for the Population Mean

## O

### observations

4.2: The Addition and Complement Rules

### outcome

4.1: Fundamental Concepts of Probability

### outcomes

4: Probability  
4.2: The Addition and Complement Rules  
4.3: Independence and Multiplication Rules

### outlier

2.1: Measures of Center of the Data

### outliers

2: Numerical Summary of Data- Center, Variation, and Location  
2.2: Measures of Spread of the Data  
2.3: Measures of Location of the Data

## P

### parameter

7.2: The Sampling Distribution of Sample Proportion

### percentile

6.2: Normal Distribution and Applications

### pie chart

1.2: Displaying Data

### point estimate

8.1: Estimating Population Proportion  
8.2: Estimating Population Mean

### population

7.2: The Sampling Distribution of Sample Proportion

### population mean

7.1: The Sampling Distribution of the Sample Mean  
8: Estimating the Value of a Parameter  
8.2: Estimating Population Mean  
9.3: Test for the Population Mean

### population parameters

7: Sampling Distributions

### population proportion

8: Estimating the Value of a Parameter  
8.1: Estimating Population Proportion  
9: Hypothesis Testing

### populations

1: Introduction to Data  
9: Hypothesis Testing

### prediction line

3.2: Least Squares Regression Line

### probabilities

6: The Normal Distribution

### probability

1: Introduction to Data  
4: Probability  
4.1: Fundamental Concepts of Probability  
4.2: The Addition and Complement Rules  
4.3: Independence and Multiplication Rules  
5: Discrete Probability Distributions  
5.1: Discrete Probability Distributions  
6.1: Properties of the Normal Distribution  
6.2: Normal Distribution and Applications

### probability distribution

5.1: Discrete Probability Distributions  
5.2: Binomial Probability Distributions  
7: Sampling Distributions

### probability of success

5.2: Binomial Probability Distributions

### properties

6: The Normal Distribution

## Q

### qualitative data

1.2: Displaying Data

### quantitative data

1: Introduction to Data

### quartiles

2.3: Measures of Location of the Data

## R

### random events

4.1: Fundamental Concepts of Probability

### random sample

7.2: The Sampling Distribution of Sample Proportion  
8.1: Estimating Population Proportion

### random variable

5.1: Discrete Probability Distributions  
5.2: Binomial Probability Distributions  
7: Sampling Distributions  
8: Estimating the Value of a Parameter

### regression analysis

3: Linear Regression and Correlation  
3.1: Scatter Diagrams and Correlation  
3.2: Least Squares Regression Line

### rejection region

9.2: Test for the Population Proportion

### relationship between variables

3.1: Scatter Diagrams and Correlation

### relative frequency

1.2: Displaying Data

### research methods

1.1: Data, Samples, and Populations

### response variable

3.1: Scatter Diagrams and Correlation

## S

### sample mean

7: Sampling Distributions  
7.1: The Sampling Distribution of the Sample Mean  
8.2: Estimating Population Mean

### sample proportion

7: Sampling Distributions  
7.2: The Sampling Distribution of Sample Proportion

## sample size

- [7.1: The Sampling Distribution of the Sample Mean](#)
- [7.2: The Sampling Distribution of Sample Proportion](#)
- [8: Estimating the Value of a Parameter](#)
- [8.1: Estimating Population Proportion](#)
- [8.2: Estimating Population Mean](#)
- [9.2: Test for the Population Proportion](#)

## sample space

- [4.1: Fundamental Concepts of Probability](#)
- [4.2: The Addition and Complement Rules](#)
- [4.3: Independence and Multiplication Rules](#)

## sample spaces

- [4: Probability](#)

## sample standard deviation

- [7: Sampling Distributions](#)

## samples

- [1: Introduction to Data](#)
- [9: Hypothesis Testing](#)

## sampling

- [1.1: Data, Samples, and Populations](#)

## sampling distribution

- [7: Sampling Distributions](#)
- [7.1: The Sampling Distribution of the Sample Mean](#)
- [7.2: The Sampling Distribution of Sample Proportion](#)

## sampling methods

- [1: Introduction to Data](#)

## scatter diagram

- [3.2: Least Squares Regression Line](#)

## scatter plot

- [3.1: Scatter Diagrams and Correlation](#)

## significance level

- [9.2: Test for the Population Proportion](#)
- [9.3: Test for the Population Mean](#)

## slope

- [3: Linear Regression and Correlation](#)
- [3.2: Least Squares Regression Line](#)

## standard deviation

- [2: Numerical Summary of Data- Center, Variation, and Location](#)
- [2.2: Measures of Spread of the Data](#)
- [2.3: Measures of Location of the Data](#)
- [5.1: Discrete Probability Distributions](#)
- [6: The Normal Distribution](#)
- [6.1: Properties of the Normal Distribution](#)
- [6.2: Normal Distribution and Applications](#)
- [7.1: The Sampling Distribution of the Sample Mean](#)
- [7.2: The Sampling Distribution of Sample Proportion](#)
- [8: Estimating the Value of a Parameter](#)
- [8.2: Estimating Population Mean](#)

## standard normal

- [6.2: Normal Distribution and Applications](#)

## standard normal distribution

- [6: The Normal Distribution](#)
- [6.1: Properties of the Normal Distribution](#)

## statistic

- [7: Sampling Distributions](#)

## statistical analysis

- [1.2: Displaying Data](#)
- [3.1: Scatter Diagrams and Correlation](#)
- [7.1: The Sampling Distribution of the Sample Mean](#)
- [8.1: Estimating Population Proportion](#)

## statistical claims

- [1.1: Data, Samples, and Populations](#)

## statistical estimation

- [7.1: The Sampling Distribution of the Sample Mean](#)

## statistical inference

- [8.2: Estimating Population Mean](#)

## statistical inferences

- [9: Hypothesis Testing](#)

## statistical significance

- [1.1: Data, Samples, and Populations](#)
- [9.1: Language of Hypothesis Testing](#)
- [9.2: Test for the Population Proportion](#)
- [9.3: Test for the Population Mean](#)

## statistical testing

- [9.1: Language of Hypothesis Testing](#)

## statistics

- [1: Introduction to Data](#)
- [1.1: Data, Samples, and Populations](#)
- [2.1: Measures of Center of the Data](#)
- [2.2: Measures of Spread of the Data](#)
- [4.1: Fundamental Concepts of Probability](#)
- [5: Discrete Probability Distributions](#)
- [6: The Normal Distribution](#)
- [6.1: Properties of the Normal Distribution](#)
- [7.2: The Sampling Distribution of Sample Proportion](#)
- [8: Estimating the Value of a Parameter](#)

## Student's t distribution

- [8: Estimating the Value of a Parameter](#)

## success

- [5.2: Binomial Probability Distributions](#)

## sum of squared errors

- [3: Linear Regression and Correlation](#)

## T

---

### test statistic

- [9.1: Language of Hypothesis Testing](#)

### theoretical

- [5: Discrete Probability Distributions](#)

### theoretical probabilities

- [4: Probability](#)

### theoretical probability

- [4.1: Fundamental Concepts of Probability](#)

### tree diagrams

- [4: Probability](#)

### trials

- [4.3: Independence and Multiplication Rules](#)

## U

---

### uncertainty

- [4.1: Fundamental Concepts of Probability](#)

## V

---

### variability

- [2.2: Measures of Spread of the Data](#)

### variables

- [1.1: Data, Samples, and Populations](#)
- [5: Discrete Probability Distributions](#)

### variance

- [2: Numerical Summary of Data- Center, Variation, and Location](#)
- [2.2: Measures of Spread of the Data](#)
- [5.1: Discrete Probability Distributions](#)
- [6.1: Properties of the Normal Distribution](#)

### Venn diagram

- [4.2: The Addition and Complement Rules](#)

### visual tools

- [1: Introduction to Data](#)



## GLOSSARY

### S

---

**Sample Word 1** | Sample Definition 1

## DETAILED LICENSING

### Overview

**Title:** Introductory Statistics STA 2023 FSW

**Webpages:** 40

**Applicable Restrictions:** Noncommercial

**All licenses found:**

- [Unknown License](#): 30% (12 pages)
- [CC BY 4.0](#): 22.5% (9 pages)
- [CC BY-SA 4.0](#): 17.5% (7 pages)
- [Undeclared](#): 10% (4 pages)
- [Public Domain](#): 10% (4 pages)
- [CC BY-NC-SA 4.0](#): 5% (2 pages)
- [CC BY-NC-SA 3.0](#): 2.5% (1 page)
- [CC BY-SA 3.0](#): 2.5% (1 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)
- [Unknown License](#): 0% (0 page)

### By Page

- [Introductory Statistics STA 2023 FSW - Undeclared](#)
  - [Front Matter - CC BY 4.0](#)
    - [TitlePage - CC BY 4.0](#)
    - [InfoPage - Public Domain](#)
    - [Table of Contents - Public Domain](#)
    - [Licensing - Public Domain](#)
  - [1: Introduction to Data - CC BY 4.0](#)
    - [1.1: Data, Samples, and Populations - Unknown License](#)
    - [1.2: Displaying Data - Unknown License](#)
  - [2: Numerical Summary of Data- Center, Variation, and Location - CC BY 4.0](#)
    - [2.1: Measures of Center of the Data - Unknown License](#)
    - [2.2: Measures of Spread of the Data - Unknown License](#)
    - [2.3: Measures of Location of the Data - Unknown License](#)
  - [3: Linear Regression and Correlation - CC BY 4.0](#)
    - [3.1: Scatter Diagrams and Correlation - CC BY-NC-SA 3.0](#)
    - [3.2: Least Squares Regression Line - Unknown License](#)
  - [4: Probability - CC BY-SA 4.0](#)
    - [4.1: Fundamental Concepts of Probability - Unknown License](#)
    - [4.2: The Addition and Complement Rules - Unknown License](#)
    - [4.3: Independence and Multiplication Rules - CC BY-SA 4.0](#)
  - [5: Discrete Probability Distributions - CC BY-SA 4.0](#)
    - [5.1: Discrete Probability Distributions - CC BY-SA 4.0](#)
    - [5.2: Binomial Probability Distributions - CC BY-SA 4.0](#)
  - [6: The Normal Distribution - CC BY 4.0](#)
    - [6.1: Properties of the Normal Distribution - CC BY-SA 3.0](#)
    - [6.2: Normal Distribution and Applications - Unknown License](#)
  - [7: Sampling Distributions - CC BY-NC-SA 4.0](#)
    - [7.1: The Sampling Distribution of the Sample](#)

- Mean - *Unknown License*
  - 7.2: The Sampling Distribution of Sample Proportion - *CC BY-NC-SA 4.0*
- 8: Estimating the Value of a Parameter - *CC BY 4.0*
  - 8.1: Estimating Population Proportion - *Unknown License*
  - 8.2: Estimating Population Mean - *Unknown License*
- 9: Hypothesis Testing - *CC BY 4.0*
  - 9.1: Language of Hypothesis Testing - *CC BY 4.0*
  - 9.2: Test for the Population Proportion - *CC BY-SA 4.0*
  - 9.3: Test for the Population Mean - *CC BY-SA 4.0*
- Back Matter - *Undeclared*
  - Index - *Public Domain*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*