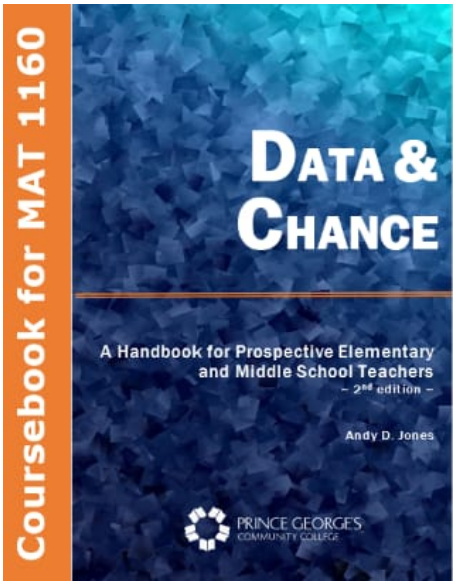
MAT 1160: DATA & CHANCE



Andy D. Jones
Prince George's Community College





Data & Chance: A Handbook for Prospective Elementary and Middle School Teachers

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

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

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



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

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

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

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

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

Title Page

Licensing

Preface

1: An Introduction to Statistics

1.1: What Is Statistics?

1.2: Populations and Samples

1.3: Levels of Measurement

1.4: Collecting Data

2: Summarizing Data Numerically

2.1: Measures of Center

2.2: Measures of Position

2.3: Measures of Spread

2.4: Changes to a Data Set

3: Visualizations of Data

3.1: Graphs of Categorical Data

3.2: Histograms, Frequency Polygons, and Ogives

3.3: Other Quantitative Graphs and Data Plots

3.4: Box-and-Whisker Plots

3.5: Normal Distribution and Its Empirical Rule

3.6: Applying a Normal Distribution

4: Correlation and Regression

4.1: Scatterplots and Correlation

4.2: Interpreting a Least Squares Regression Model

4.3: Building a Least Squares Model

5: An Introduction to Probability

5.1: Theoretical Probability

5.2: Empirical Probability and the Law of Large Numbers

5.3: Using Sets and Venn Diagrams to Represent Events

5.4: Contingency Tables, Conditional Probability, and Independence

5.5: Odds

6: Probability Rules and Expectation

6.1: Fundamental Properties of Probability

6.2: The Additive Rules of Probability

6.3: The Multiplicative Rules of Probability

6.4: Tree Diagrams and Multistage Probability

6.5: Expectation Involving Chance

7: Counting Techniques

7.1: The Fundamental Counting Principle

7.2: Permutations and Combinations

7.3: Probability Using Counting Techniques

8: Appendix- Answers and Free Graphing Tools

8.1: Answers to Section Review Exercises

8.2: Free Online Graphing Tools

Index

Glossary

Detailed Licensing

Detailed Licensing

This page titled MAT 1160: Data & Chance is shared under a license and was authored, remixed, and/or curated by Andy D. Jones.

Licensing

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

This page titled Licensing is shared under a license and was authored, remixed, and/or curated by LibreTexts.

Preface

Data & Chance: A Handbook for Prospective Elementary and Middle School Teachers

Second Edition, 2026

Andy D. Jones

Data & Chance: A Handbook for Prospective Elementary and Middle School Teachers was developed as an open educational resource (OER) for MAT 1160 *Elements of Probability and Statistics* at Prince George's Community College. It is available online and may be downloaded as a PDF. Users may print or adapt it for non-commercial purposes under the CC BY-NC-SA Creative Commons license.

MAT 1160 is the third course in the mathematics content sequence required for candidates completing the Associate of Arts in Teaching (A.A.T.) in Elementary Education at Prince George's Community College. *Data & Chance* supports the course outcomes for this 4-credit probability and statistics course designed specifically for prospective elementary and middle school teachers.

This handbook is not intended to replace a general education probability and statistics textbook. Instead, it focuses on concepts essential for teaching mathematics at the elementary and middle school levels. It also includes selected topics that extend beyond the primary grades curriculum -- such as interpreting assessment measures, understanding statistics commonly used education courses, and developing informed perspectives about data and chance as an informed consumer of information. It includes minimal topics on inferential statistics.

Data & Chance is designed as a content support resource to accompany an activity-based, inquiry-oriented course. Although the handbook does not include activities, experiments, simulations, or labs, it assumes that such experiences will be provided through other course materials.

The writing style is conversational, and formal mathematical theory is introduced through examples rather than rigorous proofs. Each section concludes with a small set of focused practice problems. When technology is needed, the handbook uses the Texas Instruments graphing calculator, with detailed step-by-step instructions provided. In addition, several free online graphing tools and calculators are identified as alternatives.

Data & Chance contains mostly original work. However, some other Open Education Resources licensed under CC BY- SA and CC BY-NC served as foundations for portions of these materials Links to these original works are shown below. A special thanks is offered to these authors and organizations.

- Topics and images from "Advanced Probability and Statistics, 2nd Edition FlexBook®" by Ellen Lawsky, Larry Ottman, Raja Almukkahal, Brenda Meery, and Danielle DeLancey, CK-12 Foundation, used under CC-BY-NC 3.0 license.
- Topics and images from *Chapter 6: Continuous Probability Distributions* of "Statistics Using Technology, 2nd edition" by Kathryn Kozak, used under CC BY- SA license.
- Topics and images from the chapters *Statistics, Describing Data*, and *Probability* from "Math in Society, Edition 2.5" by David Lippman, used under CC BY- SA 3.0 license.
- Other images and sources in the public domain are cited within the document.



This page titled Preface is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

1: AN INTRODUCTION TO STATISTICS

1.1: What Is Statistics?
1.2: Populations and Samples
1.3: Levels of Measurement
1.4: Collecting Data

This page titled 1: An Introduction to Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

1.1: What Is Statistics?

Learning Objectives

- Define statistics and key measures of central tendency and variation.
- Distinguish statistical and non-statistical questions.
- Classify data types as categorical, quantitative discrete, or quantitative continuous.

Introduction

You are exposed to statistics regularly.

- If you are a sports fan, then you may know the game statistics for your favorite player.
- If you are interested in politics, then you interpret polls to see how people feel about certain issues or candidates.
- If you are an environmentalist, then you may research arsenic levels in the water of a town or analyze global temperatures.
- If you are in the business field, then you may track the monthly sales of a store or use quality control processes to monitor the number of defective parts manufactured.
- If you are in the health profession, then you may look at how successful a procedure is or the percentage of people who have a certain disease.
- If you are a teacher, then you may collect data to summarize how well students have performed, use results of standardized tests to make informed decisions about curricula, or interpret the results of educational studies found in research journals.

Perhaps even more importantly for you as a future educator is that you will be responsible for laying the foundation of statistical reasoning for your students no matter what field or profession they plan to pursue. Even if your future students do not use statistics in their jobs, it is important that every educated citizen has some degree of statistical literacy to survive in a world that bombards us with data.

What is Statistics?

Statistics is the study of how to collect, organize, analyze, and use information collected from a group to make a decision. This information makes up a set of data. There are two branches of statistics. One is called descriptive statistics, which is where we merely collect, organize, analyze, and interpret raw data that have been collected. The other is called inferential statistics, which is where we generalize from collected data to a much larger group. We use data from a smaller group to infer something about about a larger group without collecting data from that larger group.

Two primary features that help describe data are central tendency and variability. Central tendency summarizes the data with a single value that is representative or typical of the entire data set. By variability, we mean the degree of difference or similarity observed across the data. If there were no variability in a set of data, we would not need statistics to study it because all values would be the same. Because statistics relies on studying a data set's central tendency and variability, we first must understand which questions are appropriate to ask and which questions can be answered with statistics.

Statistical Questions

An initial task in statistics is asking a question that takes variability into account. Beginning around 6th grade, students identify and pose questions that can be answered by data that vary. Students must learn to distinguish between statistical questions and questions that are not statistical. A statistical question is one that can be answered by collecting data and where we expect there will be variability in that data.

A statistical question is different from a question that anticipates a deterministic answer. For example, "*How many minutes do 6th grade students spend on homework each week?*" is a statistical question. We would answer this question by collecting data from 6th graders, and we expect that not all 6th grade students spend the same amount of time on homework (meaning there

will be variability in the data). On the other hand, "*How much time did Juana spend on homework last night?*" is not a statistical question because it has a single answer and is not answered by collecting data that vary.

Example 1.1.1

Explain whether the following is a statistical question: *How tall are you?*

Solution

To answer this question, we just have to record one height. There will be no variability in the data, which means that it is not a statistical question.

Example 1.1.2

Explain whether the following is a statistical question: *What is the typical height of the players on your favorite soccer team?*

Solution

The answer to this question we would need a list of the heights of the players on the team. We expect that there will be variability in the data we collect because not everyone will be the same height. Therefore, this is a statistical question.

Example 1.1.3

Explain whether the following is a statistical question: *How do students travel to school?*

Solution

To answer this question, we would have to collect data about how students travel to school: by car, bike, or bus, for example. Since we expect the students' answers to vary, this is an example of a statistical question.

Example 1.1.4

Explain whether the following is a statistical question: *What was the temperature on the bank sign today at noon?*

Solution

The answer to this question is a single temperature; we do not have to collect data with variability to answer the question. Therefore, it is not a statistical question.

You Try It 1.1.A

Identify whether the question in each scenario is a statistical question or not a statistical question.

a. A student wants to know how many pencils are currently in his backpack and counts them.
b. A coach wonders, "*How long does it take runners on the track team to complete a mile*?" and records times from all team members.
c. A school administrator analyzes attendance records to see how many absences students in the school accumulate over the year.

Answer available at the end of this section.

Types of Data

The information collected to answer a statistical question is a set of data. (The word *data* is plural. A single value from data is often referred to as a *datum*.) There are two broad types of data – categorical data and quantitative data.

Categorical Data

When the information can be neatly placed into well-defined groups, or categories, we have categorical data. These data may also be called qualitative because they describe a quality about a set of individuals. Usually, categorical data involve a word, name, or label that describes a characteristic of the individual. For example, data that answers the question "*How do students travel to school?*" is categorical because the information collected would be something like car, bike, bus, walk, and so on.

Quantitative Data

On the other hand, some statistical questions can be answered by counting or measuring some characteristic. These types of data are called quantitative data. Data values are numerical. We can perform meaningful arithmetic calculations. For this reason, you may also see quantitative variables referred to as numerical data. To answer the statistical question "*What is the typical height of the players on your favorite soccer team?*", we measure their heights in inches or centimeters and then do some calculation like finding an average. Quantitative data always have units associated with them.

Example 1.1.5

Are the data being collected in this scenario categorical or quantitative?

A survey asks the name of the favorite movie that moviegoers saw in a movie theater. When we conduct such a survey, the responses would look like: *Finding Nemo, The Hulk*, or *Terminator 3: Rise of the Machines*.

Solution

We might count the number of people who give each answer, but the answers themselves do not have numerical values. We cannot perform computations with an answer like *Finding Nemo*. Favorite movie is a categorical variable.

Example 1.1.6

Are the data being collected in this scenario categorical or quantitative?

A survey asks the number of movies moviegoers have seen in a movie theater in the past 12 months (0, 1, 2, 3, 4, ...).

Solution

The number of movies seen is a quantitative variable because the values are the result of counting the number of movies each person saw.

Sometimes, determining whether data are categorical or quantitative can be a bit trickier.

Example 1.1.7

Are the data being collected in this scenario categorical or quantitative?

Suppose we gather respondents' ZIP codes in a survey to track their geographical location.

Solution

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 code is really a categorical variable. We don't measure or count anything with ZIP codes. There is no unit.

Example 1.1.8

Are the data being collected in this scenario categorical or quantitative?

A survey about the movie people recently attended includes the question "*How would you rate the movie you just saw?*" with these possible answers:

1 - it was awful 2 - it was just OK 3 - I liked it 4 - it was great 5 - best movie ever

Solution

Again, there are numbers associated with the responses, but we can't really do any calculations with them. A movie that rates a 4 is not necessarily twice as good as a movie that rates a 2, whatever that means. If two people see the movie and one of them thinks it stinks and the other thinks it's the best ever, it doesn't necessarily make sense to say that "on average they liked it." Instead, the responses place attendees into categories. The variable is categorical.

You Try It 1.1.*B*

A transportation analyst observes cars at an intersection and records

a. the *type of vehicle* (sedan, SUV, truck, motorcycle),
b. the *color* of the vehicle, and
c. the *time* (in seconds) each vehicle waits at the red light.

Identify whether each data set is categorical or quantitative.

Answer available at the end of this section.

Discrete vs. Continuous

Quantitative data may be further classified as either *discrete* or *continuous*.

Discrete Data

Discrete data are values that usually result from counting. These data can take on only certain limited numerical values. The number of siblings each person has is a discrete variable. You either have 0, 1, 2, 3, and so on brothers or sisters. There is no in-between. You are counting, and so this type of data is quantitative discrete data.

Continuous Data

Continuous data are values that result from measuring and include values other than whole numbers – such as fractions, decimals, or irrational numbers. This kind of data includes values can be written more and more precisely so that there are an infinite number of values within an interval. Continuous data are often the results of measuring things like lengths of pencils, weights of babies, or times to run a marathon.

Example 1.1.9

Suppose you collect data about cell phone usage. You record the number of phone calls received each day of the past week as well as the length of each of the calls. Are these data discrete or continuous?

Solution

The data about the *number of phone calls you receive each day* is quantitative discrete because the values might look something like 0, 1, 2, 3, or even larger.

The data about the *length of the phone calls* is quantitative continuous because the values might look something like 2 minutes, 2.5 minutes, 2.52 minutes, or even 2.52364 minutes. The length of a phone call can take on an uncountable, infinite number of values greater than 0 minutes.

You Try It 1.1.C

In each scenario, identify whether the data values being recorded are discrete or continuous.

a. A researcher measures the time (in seconds) it takes for a chemical reaction to complete.
b. A school administrator counts the number of students who arrive late each morning.
c. A meteorologist measures the amount of rainfall (in inches) that falls each day during a storm system.

Answer available at the end of this section.

Section Summary

Statistics is the study of how to collect, organize, analyze, and use information called data. Data that vary can be used to answer a statistical question. Data sets can be classified as categorical or quantitative. Categorical data describe some characteristic or group to which an individual belongs. Quantitative variables are numerical data that can also be further classified as discrete (countable) or continuous (measurable.)

Section 1.1 Review Questions

1. Tell whether each question is a statistical question or not a statistical question.
   a. How many teeth does your mother have?
   b. How many days are in August?
   c. What brand of toothpaste do you use?
   d. What brand of toothpaste is most popular among your classmates?
   e. What proportion of sixth graders collect baseball cards?
   f. What types of collections do your classmates have?
   g. What is your favorite color?
   h. How old is your professor?
   i. What is the favorite subject of your classmates?
   j. How many brothers does your best friend have?
   k. What is the most commonly played musical instrument in your class?

2. Tell whether the data collected in each scenario is best classified as *categorical* or *quantitative*.
   a. name of cell phone service provider
   b. number of text messages sent per day
   c. whether or not an individual has an Instagram account
   d. how many followers each individual has for their Instagram accounts
   e. area code
   f. eye color of a group of people
   g. daily high temperature of a city over several weeks

h. the class scores on a 10-question quiz (0 1, 2, 3, ... 10)
i. the class grades on a 10-question quiz (A, B, C, D, F)

3. Tell whether the quantitative data collected in each scenario is best classified as *discrete* or *continuous*.
   a. distance from home to the nearest grocery store for students in your class
   b. number of pairs of shoes students own
   c. weights of sumo wrestlers
   d. GPAs of students in your class
   e. number of pitches made by the starting pitcher in games this season
   f. number of heads when tossing three pennies
   g. shoe sizes sold during a shoe sale

Answers

You Try It $1.1.A$

a. not statistical
b. statistical
c. statistical

You Try It $1.1.B$

a. categorical
b. categorical
c. quantitative

You Try It $1.1.C$

a. continuous
b. discrete
c. continuous

This page titled 1.1: What Is Statistics? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

1.2: Populations and Samples

Learning Objectives

- Identify key components of a statistical study.
- Differentiate populations and samples.
- Differentiate parameters and statistics.
- Explain sampling variability and error.

Introduction

To understand how to calculate descriptive statistics to make inferences about the world around us, there are a few definitions that we need to understand. Many of the terms used in statistics also have non-statistical meanings. That is, in statistics, some terms have slightly different definitions than when we use those terms in everyday language or in other branches of mathematics. It is important that you notice the subtle differences and use the statistical definitions when appropriate.

Population vs. Sample

In any statistical study, the first task is to determine **who** is being studied and **what** is being measured. These two questions form the foundation of every well-designed study. The *who* refers to the group of **individuals** (or units) under consideration—these may be people, animals, manufactured items, or any units of interest. The *what* refers to the **variable**, the specific characteristic or measurement recorded for each individual.

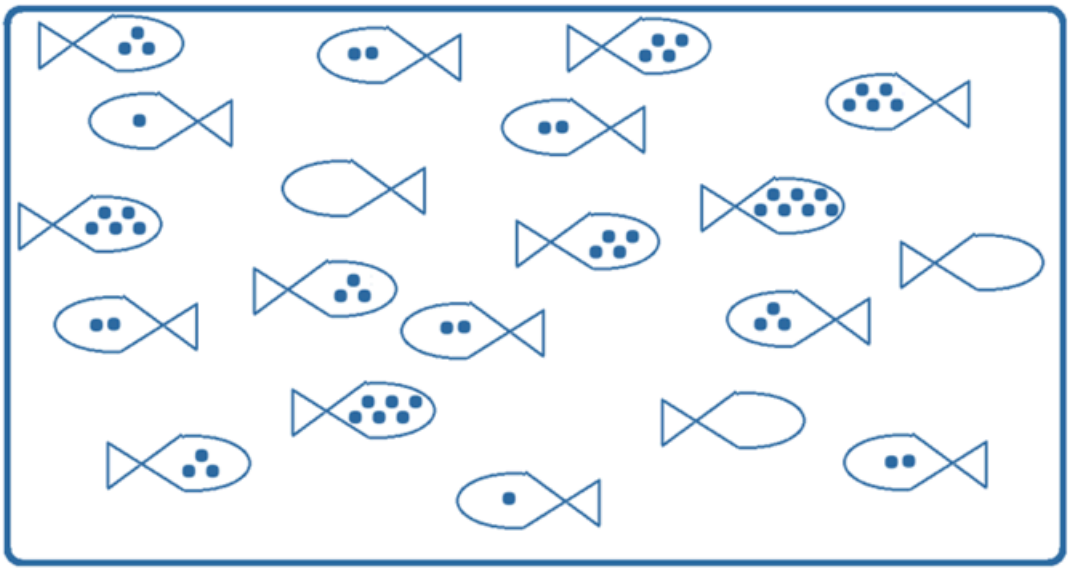
The **population** is the entire group of individuals about whom information is desired. A clear population statement must always identify both *who* is being measured and *what* is being measured. Without both components, the purpose of the study is incomplete. For example, stating that a researcher will "collect data from the senators in the U.S. Congress" does not tell the reader what attribute is being studied. Are we interested in age, years of service, educational background, voting behavior, or something else? A population description must specify the individuals and the variable to give a complete picture of the study's intent.

In some cases, the population is small enough that measuring every individual is practical. When data are collected from the entire population, the study is called a **census**. For instance, determining the current ages of all 100 sitting U.S. senators is straightforward because the population is limited and accessible. However, if the goal were to determine the age at which every senator first took office throughout the entire history of Congress, the population is still finite but requires more extensive data collection.

Many populations, however, are too large or too difficult to measure in their entirety. In such situations, researchers select a smaller group from the population. This subset is called a **sample**. A sample serves as a manageable representation of the population—ideally similar in its characteristics but far easier to measure. By studying the sample, researchers draw conclusions about the larger population.

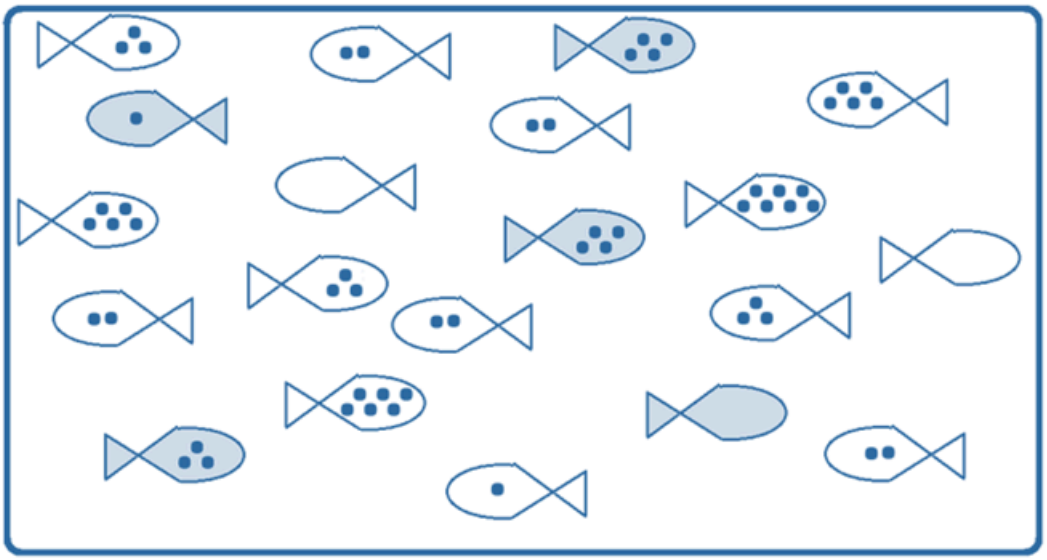
For a visual illustration, *Figure 1* shows a fish tank containing several fish -- the population. The characteristic being studied - the variable - is the number of spots on each fish. Here, the 20 individual fish are the *who* and the number of spots on each of the fish is the *what*.

Figure 1: Population of Fish in Fish Tank



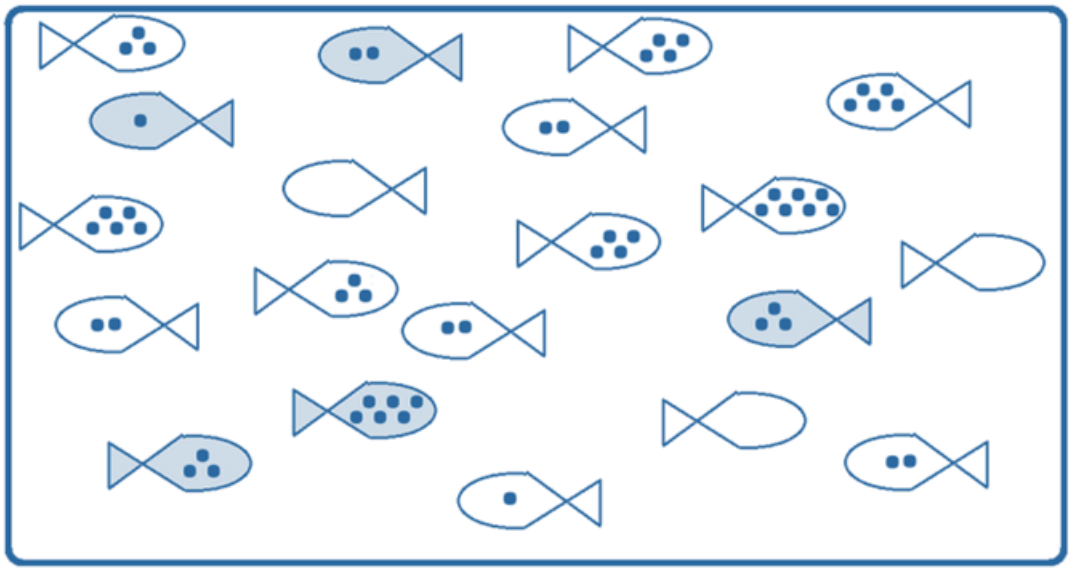
To collect a sample of the fish from the tank, perhaps we use a net to select 5 fish at random. The fish in Sample A are shaded in *Figure 2*. Data are being obtained from the 5 fish – not all 20 fish – in hopes that the subgroup of 5 fish resembles the larger group.

Figure 2: Sample A of 5 Fish from the Fish Tank



Realize that if we repeat this sampling process, we may get the same 5 fish again, some of the same 5 fish again and some different fish, or all different fish. *Figure 3* shows Sample B, a different sample of 5 fish taken from the original population of 20 fish:

Figure 3: Sample B of 5 Fish from the Fish Tank



Two of the same fish were caught in both samples, but the samples are quite different. That is, when you take a sample you just get a "snapshot" of the population and this snapshot may be different each time. The idea is to try to collect a sample that gives an accurate overall picture, but we never know for sure if the snapshot does that. Unlike previous mathematics classes where there was always one right answer, in statistics there can be many answers, and you don't know which are right. We use probability later in this course to estimate our likelihood of being right.

In statistics, it is convention to always use the letters N and n for the size of the population and the size of the sample, respectively. For the previous fish tank illustration, there were 20 fish in the population under study so $N = 20$. To study the fish in the fish tank, we took samples of size 5 fish so $n = 5$.

How you collect your sample can determine how accurate the results of your study are as well. We hope that the sample resembles the population as closely as possible. There are many ways to collect samples. Some of them create better samples than others. More on that will be discussed later.

Why Sample?

In the last illustration, the population of fish was quite small (only 20 fish), and it was manageable to conduct a census of the fish in the fish tank. Most of the time, however, it is extremely difficult or very costly to collect all the information about an entire population. For a study of tortoises in the Galapagos Islands, it would be very difficult and perhaps even destructive to search every square meter of the habitat to be sure that you reached every tortoise. A battery manufacturer would not want to test every battery it produces to determine the average lifetime of its batteries because it would have none left to sell. It is very expensive to get accurate and complete information about all the residents of the United States to help effectively address the needs of a changing population. For that reason, the U.S. Government conducts a census only once every ten years.

Example 1.2.1

To find the most common manufacturer, a campus security officer records the manufacturer of every 5th car that enters the college parking lot until data about 40 cars has been collected. Identify the *individual*, the *variable*, the *population*, and the *sample*, n, and N, where possible. Also tell if the data collected in the scenario is *categorical* or *quantitative*.

Solution

- Individual: a car that enters the college parking lot
- Variable: the name of the car manufacturer
- Population: all cars entering the college parking lot
- Sample: every 5th car entering the college parking lot

- $n = 40$ because data will be collected from 40 cars
- N is unknown. We were not told how many cars will enter the college parking lot.
- Categorical data because the name of a car manufacturer in being recorded

Example 1.2.2

A biologist planted 500 Variety X tomato plants and gave them a new plant food. Now, the biologist wants to estimate the average height of the plants. She randomly selects 40 of the tomato plants and measures the heights.

Identify the *individual*, the *variable*, the *population*, and the *sample,* n, and N, where possible. Also tell if the data collected in the scenario is *categorical* or *quantitative*.

Solution

- Individual: a Variety X tomato plant given the new plant food
- Variable: the height
- Population: all the Variety X tomato plants that were planted and given the new food
- Sample: only the plants that the biologist measured heights for
- $n = 40$ because data was collected from 40 plants in the sample.
- $N = 500$ because the entire group under study contains 500 plants
- Quantitative continuous data because the data being collected consists of numbers that come from measuring height in inches or centimeters.

You Try It 1.2.A

A school administrator wants to estimate the average number of hours students at Lincoln High School spend on homework each week. She randomly selects 60 students and asks each one how many hours they spent on homework during the past week.

Identify the *individual*, the *variable*, the *population*, the *sample,* n and N, where possible. Also tell if the data collected in the scenario is *categorical* or *quantitative*.

Answer available at the end of this section.

Parameter vs. Statistic

Once data have been collected—either through a census or from a sample—the next step is typically to summarize the information by performing calculations. Depending on the study, a researcher may wish to determine the proportion of residents who support a mayoral candidate, the average growth of a plant treated with a new fertilizer, or the variability in a set of test scores. When these calculations are based on every unit in the population, the resulting numerical value is called a **parameter**. Parameters are usually represented with Greek letters like ρ, μ, and σ.

As noted earlier, collecting data from an entire population is often impractical or impossible. Although the parameter is the value we ultimately wish to know, it cannot be computed directly when the full population is not observed. Instead, researchers rely on a numerical value computed from a sample, known as a **statistic**, to estimate the parameter. Because samples differ from one another, the statistic will vary from sample to sample. It provides an estimate of the parameter, but its accuracy cannot be guaranteed. Statistics are typically represented with Latin letters and symbols (like $\hat{p}$, $\bar{x}$, or s).

Example 1.2.3

Use *Figure 1* to find the average number of spots per fish in the fish tank. Is this average a parameter or a statistic?

Solution

Conduct a census. There are $N = 20$ data values in this population so find the total number of spots and divide by 20:

$$\begin{aligned} &= \frac{\begin{array}{c}(3+1+5+2+3+2+0+3+2+6 \\ +4+2+4+1+7+3+0+5+0+2)\end{array}}{20} \\ &= \frac{55}{20} \\ &= 2.75 \end{aligned}$$

This value is a parameter because all values from the population were used in the calculation of the average. We can write the average as $\mu = 2.75$ spots.

Example 1.2.4

Use *Figure 2* and *Figure 3* to find the average number of spots per fish in Sample A and the average number of spots per fish in Sample B. Are these averages parameters or statistics?

Solution

There are 5 data values in Sample A so find the total number of spots and divide by $n = 5$:

$$\frac{1+3+4+4+0}{5} = \frac{12}{5} = 2.4$$

We can write the average for Sample A as $\bar{x} = 2.4$ spots.

There are 5 data values in Sample B so find the total number of spots and divide by $n = 5$:

$$\frac{1+3+2+6+3}{5} = \frac{15}{5} = 3$$

We can write the average for Sample B as $\bar{x} = 3$ spots.

Both of the calculated values in Example 1.2.4 are statistics because only some of the fish from the population were used to estimate the average number of spots for the entire population. Note that the two samples produced different statistics, but both are reasonable estimates of the population parameter – one is a bit lower and one is a bit higher and neither are exactly correct.

You Try It 1.2.B

A health researcher wants to estimate the proportion of adults in a city who walk at least 30 minutes per day. She collects data from a random sample of 400 adults and finds that 62% meet this guideline.

Classify each value:

a. The percentage of all adults in the city who walk at least 30 minutes per day.
b. The percentage of the 400 sampled adults who walk at least 30 minutes per day.
c. The average age of all adults in the city.
d. The average age of the 400 sampled adults.

Answer available at the end of this section.

Sampling Error

Researchers must recognize that statistics calculated from a sample always carry some degree of uncertainty. This uncertainty is unavoidable unless the entire population is measured through a census. Because samples vary by chance, results will differ from

one sample to another.

A statistician can report the estimated average weight of a blue crab in two ways: as a *point estimate* (for example, 0.35 pounds) or as an interval estimate. A typical interval estimate statement might be: "*I am fairly confident that the average weight of a blue crab lies between 0.3 pounds and 0.4 pounds.*" This range reflects the natural variability that arises from sampling—not an error in data collection or analysis.

The difference between the true population parameter and the sample statistic is known as **sampling error**. These concepts of estimates and errors will be explored further in a later chapter.

Section Summary

In statistics, the entire group of data being studied is called the population. The entities (people, animals, or things) in the population are called individuals. The characteristics of interest to us are called variables.

Because of the difficulties of obtaining information about all individuals in a population, it is common to use a small, representative subset of the population called a sample. A calculated value from population data (percentage of blue crabs in the Chesapeake Bay that are female, average weight of all blue crabs, etc.) is called a parameter. We can estimate a parameter by collecting a sample and calculating a statistic.

Whenever data from samples are used instead of from a census, we must accept that our results will vary from sample to sample due to sampling error. These results are merely estimates of the parameter value.

Section 1.2 Review Questions

1. In each of the following situations, identify the *individual, variable, population*, and *sample*. Also tell if the variable is *categorical* or *quantitative*.
   a. Sweet Tooth Candy Company makes chocolate candy bars. A quality control worker weighs every 100th candy bar the company makes to ensure that it is very close to the published weight.
   b. Lucio wants to know whether the hot food he serves in his restaurant is within a safe range of temperatures. He randomly selects 10 entrees and measures their temperatures just before he serves them to his customers.
   c. Suppose you want to know how Maryland workers age 18 or older travel to work. To estimate the percentage of people who use the different modes of travel, you take a sample containing 500 Maryland workers age 18 or older.
   d. You want to find the average cholesterol level of patients two days after they had a heart attack. To estimate the average, a doctor collects data from 8 heart patients in a large hospital who had a heart attack two days ago.

2. Decide whether each underlined value is a *parameter* or *statistic*.
   a. A study of 600 adults using public rest rooms in Atlanta, Chicago, New York City, and San Francisco found that **23%** did not wash their hands before exiting.
   b. **32%** of U.S. kindergarten teachers say that knowing the alphabet is an essential skill.
   c. At a college, **90%** of the members of the Board of Trustees approved the contract of the new president.
   d. The 2182 students who accepted admission offers to Northwestern University in 2009 have an average SAT score of **1442.**
   e. A researcher wants to know the average height of women aged 20 years or older. The researcher calculates an average height of **63.9** inches from measuring 45 women.
   f. Ty Cobb is one of major league baseball's greatest hitters of all time and had a career batting average of **0.366.**

Answers

You Try It $1.2.A$

- Individual: one student who was asked about weekly homework hours
- Variable: the number of homework hours each student reports
- Population: the number of homework hours completed by each of the students at Lincoln High School
- Sample: the number of homework hours reported by each of the 60 students selected and surveyed
- $n = 60$
- N is unknown. It would be the number of students at Lincoln High School.
- Quantitiative data

You Try It $1.2.B$

a. parameter
b. statistic
c. parameter
d. statistic

This page titled 1.2: Populations and Samples is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

1.3: Levels of Measurement

Learning Objectives

- Identify levels of measurement for variables as nominal, ordinal, interval, or ratio.

Introduction

Not all statistical methods and graphs can be used for all data. The levels at which data are collected help us determine what statistical technique is appropriate to use. For example, if the level of measurement of your data is nominal (the least precise and informative variable), you can use mode to summarize your variable, but not a median nor mean. If your data are at an interval or ratio level, you can use the mean, median, and mode to summarize the data. This lesson is an overview of the basic considerations involved with the levels at which you can measure a variable.

Levels of Measurement

While data can be generally classified as either categorical or quantitative, sometimes researchers and statisticians use a more detailed distinction when examining the information that is collected. The **level of measurement** refers to the way a variable's values can be categorized, ordered, or quantified. It tells you what kind of information the data provide so you know what statistical procedures are appropriate. Every variable in a study has a specific level of measurement, and understanding that level is essential for choosing correct summaries, graphs, and analyses.

This widely accepted classification includes four levels of measurement: *nominal, ordinal, interval*, and *ratio*. Each level is addressed below briefly.

Nominal Level

Nominal data represent the lowest level of measurement, where the values of a variable function solely as names, labels, or categories. These categories have no inherent order and no meaningful numerical relationship. Because the values are not numerical quantities, no arithmetic operations—such as addition, subtraction, or averaging—can be performed on nominal data.

Common examples of nominal variables include gender, car model, ethnicity, and race. Another example is political affiliation, where voters may be classified as Democrat, Republican, Libertarian, Green, or other parties. In each case, the categories identify group membership but do not imply any ranking or magnitude.

Nominal data are essential for classification and grouping, but their lack of order and numerical meaning limits the types of statistical analyses that can be applied.

Ordinal Level

Ordinal data represent a level of measurement in which the values of a variable can be placed in a meaningful order, but the differences between those values are not numerically precise. The categories indicate rank or position, yet the spacing between ranks is not consistent or measurable. Because the values do not represent equal numerical intervals, arithmetic operations cannot be performed on ordinal data.

Variables measured at the ordinal level include

- class rankings (freshmen, sophomore, junior, senior),
- levels of satisfaction (such as poor, fair, good, excellent), and
- Likert scales (strongly disagree → disagree → neutral → agree → strongly agree).

Another example of data measured at the ordinal level would be recording drink sizes sold at a movie theater as *small*, *medium*,

and *large*. These categories clearly reflect an order, but the difference between “small” and “medium” is not necessarily the same as the difference between “medium” and “large,” and none of the values represent a precise numerical quantity.

Ordinal data allow researchers to compare categories in terms of “greater than” or “less than,” but they do not support meaningful calculations involving differences or averages.

Interval Level

Interval data represent a level of measurement in which the values of a variable are numerical, ordered, and have meaningful, consistent differences between them. Because the spacing between values is uniform, subtraction is meaningful and differences can be interpreted directly.

A common example of interval data is temperature measured in degrees Fahrenheit or Celsius. The difference between 70°F and 75°F is the same as the difference between 80°F and 85°F. Another example is calendar years, where the difference between years is meaningful and consistent.

However, interval data do not have a true zero, meaning the zero point does not represent the complete absence of the quantity being measured. For instance, 0°F does not indicate “no temperature,” and the year 0 does not represent the absence of time. In both cases, the value 0 is arbitrary and just a point on a scale. Because interval scales lack a true zero, ratios are not meaningful. It would not make sense to claim that 60°F is “twice as hot” as 30°F.

Interval data support addition and subtraction, but not multiplication or division, because ratios do not carry meaningful interpretations.

Ratio Level

Ratio data represent the highest level of measurement, where values are numerical, ordered, and have consistent intervals between them. Most importantly, ratio data include a true zero, indicating the complete absence of the quantity being measured. Because of this true zero, ratios are meaningful, and statements such as “twice as much” or “half as much” are valid.

Examples of ratio variables include height, weight, time, distance, and number of hours studied. For instance, a weight of 0 pounds represents no weight, and a student who studies 10 hours has studied twice as long as a student who studies 5 hours. Both differences and ratios carry meaningful interpretations.

Ratio data allow all arithmetic operations—addition, subtraction, multiplication, and division—making them the most flexible and informative type of quantitative measurement.

Determining the Level of Measurement

Many students struggle to determine whether a variable is measured at the nominal, ordinal, interval, or ratio level. A helpful starting point is to identify whether the variable is categorical or quantitative.

If the variable is categorical, then it is generally either nominal or ordinal. Ask yourself whether the categories can be placed in a meaningful order. If the categories have no natural ordering, the variable is nominal. If the categories can be ordered, the variable is ordinal.

If the variable is quantitative, then it is typically either interval or ratio. The key distinction is whether the scale includes a true zero. For ratio-level data, a value of 0 represents the complete absence of the quantity being measured. If the variable has an absolute zero, it is ratio; if it does not, it may be interval.

A word of caution: Ordinal variables are sometimes displayed using numbers—for example, “5” for strongly agree and “1” for strongly disagree. These numbers are not inherent numerical values; they are simply labels assigned to ordered categories. Although many people compute averages or other statistics with such data, these operations are not technically appropriate. When numbers appear, make sure they represent true numerical quantities rather than arbitrary codes.

While categorical variables are usually nominal or ordinal, exceptions exist. Consider the variable year of birth for students in a class (e.g., 1970, 1994, 2004). Years of birth are categorical in the sense that you would not perform arithmetic with them and they

do not represent measured units. However, they are measured at the interval level because the difference between consecutive years is meaningful and consistent. The difference between 1970 and 1971 is the same as the difference between 1971 and 1972. This is an example of a variable that is categorical in form but measured at the interval level.

Let's use an example to show how you could collect data at different levels of measurement from the same population.

Example 1.3.1

Assume a guidance counselor wants to collect data about the students in her school.

Nominal Level: Students' gender, race, political opinions, or the town in which they live are measured at the *nominal level* because the values have no natural order.

Ordinal Level: Students' year in school (freshman, sophomore, junior, senior) have a rank and thus are at the *ordinal level*.

Interval Level: Students' SAT math scores are numerical, they can be ordered, and the difference between any two scores is meaningful. The difference between a score of 520 and 540 reflects the same change in measured performance as the difference between 660 and 680. However, SAT Math scores do not have a true zero. A score of 200—the lowest possible score—does not represent "no mathematical ability." It is simply the bottom of the reporting scale. You cannot say that a student who scores 600 is twice as good at math as a student who scores 300, so ratios are meaningless and the data is at the *interval level*.

Ratio Level: Student's age, height, weight, and GPA each have an absolute zero that has real meaning. While the data are numerical, can be ordered, and have differences that are meaningful, ratios also can be formed. Someone who is 18 years old is twice as old as a 9-year-old. These variables would measure at the *ratio level*.

The table below summarizes the four levels of measurement and provides an example of data at each level.

Levels of Measurement

Level	Description	What You Can Do	What You Cannot Do	Examples
Nominal	Categories with no order; labels only	Count frequencies; identify modes	No ordering; no arithmetic	Gender, car model, political affiliation
Ordinal	Categories with a meaningful order, but unequal spacing	Rank values; compare "greater/less"	No meaningful differences; no arithmetic	Satisfaction ratings, class rank, drink size (small, medium, large)
Interval	Numerical values with equal intervals, but no true zero	Add and subtract; interpret differences	No ratios ("twice as much" is invalid)	Temperature (°F, °C), calendar years
Ratio	Numerical values with equal intervals and a true zero	All arithmetic; meaningful ratios	—	Height, weight, time, distance

Example 1.3.1

Jacob is studying famous tourist attractions around the world for a school project. For each attraction, he collects some data.

I. Jacob records whether the tourist attraction is in an area that is urban, rural, dessert, or rain forest. The variable is being measured at the *nominal level*. The attractions can be categorized, but there is no natural order to the categories.
II. Jacob records elevations for each attraction as high elevation, medium elevation, or low elevation. This variable is measured at the *ordinal level*. The categories have a natural order.
III. Jacob records the latitude of each attraction. This variable is measured at the *interval level*. A value of 0 means the attraction lies on the equator – not that the attraction has no latitude. The value of 0 is an arbitrary starting point assigned to the equator. An attraction with 60°N latitude does not have twice as much latitude as an attraction with 30°N.

IV. Jacob records the number of miles the attraction is located from Washington, DC. The variable is measured at the *ratio level*. If one attraction is 500 miles from Washington, DC, and a second one is 1000 miles away, the second one is twice as far away. Ratios make sense.

You Try It $1.3.A$

Identify the level of measurement:

Military title (*private, sergeant, lieutenant, captain, major, colonel, general)*

Answer available at the end of this section.

You Try It $1.3.B$

Identify the level of measurement:

clothing purchased (*hat, shirt, shoes, pants, socks*)

Answer available at the end of this section.

You Try It $1.3.C$

Identify the level of measurement:

number of of siblings each of your classmates has (0, 1, 2, 3, etc.)

Answer available at the end of this section.

You Try It $1.3.D$

Identify the level of measurement:

IQ of each of your family members (100, 115, 85, 145, etc.)

Answer available at the end of this section.

Section Summary

Variables can be measured at different levels depending on the nature of the variable and the amount of information the measurement provides. A widely used framework classifies measurements into four levels. Nominal data consist of names or categories used only for identification or classification. Ordinal data also use categories, but these categories have a meaningful order or ranking. Interval measurements provide ordered numerical values with equal spacing between them, allowing meaningful differences to be interpreted but no true zero point. Finally, ratio measurements include all the properties of interval data but also feature a true zero point, making ratios between values meaningful. Understanding these levels helps determine which statistical methods and calculations are appropriate for a given variable.

Section 1.3 Review Questions

1. Identify the level of measurement for each variable described below.
   a. amount of time you watch television on a given day
   b. type of television show (comedy, drama, cartoon, sports, *etc.*)

c. Olympic medal types (gold, silver, bronze)
d. low temperatures in Fairbanks, Alaska each day of the year
e. how much you like spinach (really like, like, it's okay, dislike, really dislike)
f. amount of time you spent on homework last night

2. Identify the level of measurement for each variable described below.
   a. Personality type (introvert, extrovert, ambivert)
   b. Level of education completed (*high school, associate's degree, bachelor's degree, master's degree*)
   c. year in which couples were married
   d. weight of kindergartners
   e. test grades recorded as A, B, C, D, or F
   f. employment status (*unemployed, part-time, retired, full-time*)

Answers

You Try It $1.3.A$

Ordinal level. There is an inherent ordering or rank in that a major is higher than a captain, which is higher than a lieutenant. However, the difference between lieutenant and captain is not necessarily the difference between captain and major.

You Try It $1.3.B$

Nominal level. Clothing type can be categorized but have no natural order.

You Try It $1.3.C$

Ratio level. A data value of 0 means that a classmate has no siblings. A classmate with 6 siblings has three times as many siblings as a classmate with two siblings.

You Try It $1.3.D$

Interval level. An IQ score is a scaled score. Difference between 85 and 100 is the same as between 100 and 115, but IQ of 0 wouldn't mean you have no intelligence.

This page titled 1.3: Levels of Measurement is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

1.4: Collecting Data

Learning Objectives

- Define bias as it relates to sampling and statistical studies.
- Evaluate sampling situations for bias.
- Identify types of bias.
- Identify basic sampling techniques.

Introduction

Before we can summarize or analyze data, we must first understand how the data are obtained—and not just any data, but good, trustworthy, representative data. In statistics, data typically come from two main sources: samples and experiments. Regardless of the source, the goal is to collect information that accurately reflects the population of interest.

When data are gathered in a way so they are representative of the population, we say the data are **unbiased**. A sample or experiment should mirror the population's characteristics as closely as possible. For example, if you want a sample of fish from a lake, you should collect fish from multiple locations rather than just one spot. Likewise, if you are testing a new painkiller for adults, your experiment should include a diverse group of participants—young and old, healthy and unhealthy, male and female, and individuals with a range of body types and backgrounds. The more your sample resembles the population, the more confident you can be that your results generalize beyond the individuals you studied.

Whether data come from a sample or an experiment, they should not be influenced by the method of collection. If the way the data are selected or gathered systematically favors certain outcomes or certain types of individuals, the resulting data are **biased.** Bias can arise from poor sampling methods, non-random selection, leading questions, or experimental procedures that unintentionally influence participants.

This section provides an overview of how bias can affect the quality of data and introduces techniques—such as randomization and thoughtful sampling design—that help minimize bias. Understanding these pitfalls is essential for producing reliable data and drawing valid conclusions from statistical studies.

Sources of Bias in Samples and Surveys

Obtaining a sample that truly reflects an entire population is often challenging. Even small errors in how a sample is selected or how information is collected can produce a non-representative sample, leading to conclusions that differ from those we would obtain from a complete census. Any systematic influence that pushes the data away from accurately reflecting the population is called **bias**. Bias can arise from the selection process, the way questions are asked, or the way responses are provided.

In statistical studies, recognizing and minimizing bias is essential for producing reliable results. Three major forms of bias commonly affect samples and surveys: selection bias, response bias, and non-response bias.

Selection Bias

Selection bias occurs when the method used to choose a sample systematically favors certain individuals or groups, causing the sample to misrepresent the population. This happens when some members of the population have a greater or lesser chance of being included than others. As a result, the conclusions drawn from the sample may differ from those we would obtain if every member of the population had been studied.

Selection bias can arise in many ways. For example, surveys that rely on people choosing to participate often attract individuals who feel strongly about the issue, while those with milder opinions may not respond. Samples drawn from easily accessible

groups—such as friends, classmates, or nearby individuals—may over-represent people who share similar backgrounds or viewpoints. Bias can also occur when the list used to select participants (called a sampling frame) excludes certain groups, such as a telephone survey that misses people without landlines.

In each case, the problem is the same: the sampling process gives some members of the population more opportunity to be selected than others. When this happens, the resulting data may not accurately reflect the population, and any conclusions based on the sample may be misleading.

Response Bias

Response bias occurs when the way a survey is presented influences how individuals in the sample answer, causing their responses to differ from their true beliefs or behaviors. Unlike selection bias—which affects who ends up in the sample—response bias affects how the selected individuals respond once they are included.

Response bias can arise when the wording of a question leads people toward a particular answer or when a question is confusing, emotionally charged, or framed in a way that encourages a specific viewpoint. In some cases, respondents may hesitate to answer honestly if the topic feels sensitive or personal, particularly when they fear their answers might be judged or traced back to them. People may also give answers they believe are socially acceptable rather than answers that reflect their actual opinions or behaviors. It can also occur when respondents struggle to remember past events accurately, especially in surveys that rely on self-reporting.

In all these situations, the responses collected do not accurately represent the individuals' true thoughts or experiences. Instead, they reflect the influence of memory limitations, question wording, social pressure, or concerns about privacy. When response bias is present, the resulting data can be misleading, even if the sample itself was selected appropriately.

Non-Response Bias

Non-response bias occurs when individuals who are selected for a sample choose not to participate or discontinue their participation. Even when the sampling method is sound, the final sample may fail to represent the population if a substantial portion of selected individuals do not respond. In many surveys, people simply do not want to take the time to participate—they may hang up on a telephone poll, discard a mail-in questionnaire, or walk past an interviewer without stopping. When this happens, the views or characteristics of non-respondents are missing from the data, and the results may be skewed toward those who were willing to participate.

Non-response bias is especially problematic because the people who opt out often differ in meaningful ways from those who respond. Their absence can distort estimates, weaken conclusions, and reduce confidence in the results. Since nearly all surveys are vulnerable to non-response, statisticians must think carefully about how to encourage participation and reduce the likelihood that important segments of the population are left out.

Example 1.4.1

In each situation, identify the potential type of bias (*selection bias, response bias*, or *non-response bias*) and explain why there may be bias.

I. A survey asks high school students if they have ever cheated on a test.

Response bias. Respondents may fear repercussions or stigma associated with cheating and answer dishonestly.

II. A radio station asks readers to phone in their choice for a vice-presidential nominee in a daily poll.

Selection bias. Respondents are self-selected and most likely are those who have strong feelings about politics one way or the other. Listeners of radio stations also tend to have similar demographics such as age, education level, and so on.

III. A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 6 students sitting in the front row their last test score.

Selection bias. Data are collected only from those who sit in the front of the class. If students who sit in the front are different in terms of grades earned from those who sit elsewhere, the data may be biased.

IV. A county administrator mails a survey printed in English to 1,000 households in a predominantly immigrant neighborhood about whether English is spoken in the household. The results of the survey show that about 85% of households in this neighborhood speak English, based on the 120 surveys that were returned.

Non-response bias. A very low response rate was seen here. We cannot be sure that the relatively few surveys returned are representative of the intended population. Perhaps those who don't speak English were not able to read and complete the survey thereby inflating the percentage of households who speak English.

V. A poll asks, “Do you support a new transportation tax, or would you prefer to see our public transportation system fall apart?”

Response bias. The poll uses a loaded question that leads a respondent towards a socially preferable answer.

Randomization: A Remedy to Eliminate Bias

Bias can create serious problems for statisticians because it leads to samples that do not accurately represent the population. One of the most effective ways to reduce bias in sampling is through **randomization**. Randomization removes human choice from the selection process and ensures that every member of the population has an equal chance of being included. When a **simple random sample** of size n is taken, all possible samples of that size are equally likely to be selected. This helps prevent systematic influences—such as convenience, personal preference, or hidden patterns—from distorting the sample.

Randomization can be carried out in many practical ways. For example, a teacher who wants to choose a student at random to present a solution at the board could place all students’ names in a box, mix them thoroughly, and draw one. More efficiently, the teacher could assign each student a number and use a calculator or computer to generate a random number corresponding to the selected student. In both cases, chance—not judgment—determines who is chosen, helping ensure that the selection is fair and unbiased. By relying on chance rather than choice, randomization helps produce samples that more accurately reflect the population. It is a foundational technique in statistics and a key safeguard against bias in sampling.

Sampling Techniques

Statisticians use several methods to select samples from a population. Some methods rely on randomization and are designed to produce representative samples, while others are prone to bias and should be used with caution. Four commonly used probability-based sampling methods are *simple random sampling*, *systematic sampling*, *stratified sampling*, and *cluster sampling*. Two additional methods—*convenience sampling* and *voluntary response sampling*—are easy to carry out but often lead to biased results.

Simple Random Sample (SRS)

A **simple random sample** (or SRS) is the most fundamental method. In a simple random sample, every member of the population has an equal chance of being selected, and every possible sample of a given size is equally likely. For example, a teacher could place all students’ names in a box, mix them thoroughly, and draw several names, or assign each student a number and use a random number generator to select the sample. Because the selection is based entirely on chance, simple random sampling is one of the best ways to avoid bias.

Systematic Sample

A **systematic sample** selects individuals at regular intervals throughout the population. For instance, a researcher surveying shoppers might choose every 10th person who enters a store after a randomly chosen first shopper.

To take a 1-in-k systematic sample, first choosing a starting individual from the first k individuals. Then, select every k^{th} individual.

Systematic sampling is easy to implement and ensures coverage across the entire population. It is not a simple random sample because not all possible combinations are equally likely. It should not be used if there are patterns in the way that the population is ordered.

Example 1.4.2

To take a sample of some students from her class of $N = 30$ students, a teacher chooses the third student on the roster and then every 6th student thereafter.

This is a 1-in-6 *systematic sample* because individuals at regular intervals (every 6th student in this case) was selected. The students selected for the sample would be #3, #9, #15, #21, and #27.

Stratified Sample

A **stratified sample** divides the population into distinct subgroups, or strata, based on shared characteristics such as age, grade level, or income. A simple random sample is then taken from each stratum. In a survey of Prince George's County high school students, we might choose to stratify by high school to be sure that the opinions of all school locations in the county are included. We could simply choose to take an SRS of size 20 students from each school in the county and then combine them to form a stratified sample.

Stratified sampling ensures representation from all key subgroups and is especially useful when the population is diverse. Usually statisticians try to make the number of individuals chosen from each strata proportional to the population. For example, if High School A is really large and High School B is really small, then more students should be sampled from High School A than from High School B.

Example 1.4.3

A new state proposal is on the ballot in November's election. The state voter registration data indicates that the electorate is made up of 39% Democrats, 37% Republicans and 24% independents. For a sample to fairly represent these different groups, the sample of 1000 people should contain 390 Democrats, 370 Republicans and 240 independents. To accomplish this, pollsters randomly select 390 people from among those voters known to be Democrats, 370 from those known to be Republicans, and 240 from those with no party affiliation.

This is *stratified sampling*. The strata are the three political affiliations – Democrat, Republican, and Independent. Some voters are randomly selected from each of the strata in proportion to their prevalence in the population. Without stratifying, there would be no guarantee that the makeup of the sample would reflect the natural variation in political affiliation in this state.

Cluster Sample

A **cluster sample** divides the population into naturally occurring groups called clusters. One or more clusters are selected at random, and all individuals within those clusters are included in the sample. For example, a citywide survey might randomly select several blocks and collect data from every household on those blocks. For a cluster sample to be effective, it is important that the clusters are alike.

The National Park Police sometimes uses cluster sampling to determine the number of people who show up for events on the National Mall. They may take a picture of a few small, representative areas of the crowd, count the individuals in just those area,

and then use those counts to estimate the total crowd in attendance.

Example 1.4.4

The PGCC Math Department wants to survey students about use of its tutoring center. Since students are already divided into math classes, the researchers randomly select 10 classes and give the survey to all the students in those classes.

This is *cluster sampling*. Each math class is a cluster. Once a class is selected at random, every student in that class is surveyed. If a class is not selected, then no one in that class is included in the survey.

Many people confuse stratified sampling and cluster sampling:

- In stratified sampling you use all the groups but only select some members from each group.
- Cluster sampling is the other way around. You use only some of the groups but select all members in each group.

Convenience and Voluntary Response Samples

While the methods above use randomization to reduce bias, two other sampling techniques—convenience sampling and voluntary response sampling—do not. **Convenience samples** contain individuals who are easiest to reach, such as surveying friends, classmates, or people nearby. **Voluntary response samples** relies on individuals choosing to participate, such as call-in polls or online opinion surveys. Both methods tend to over-represent people with strong opinions or similar backgrounds, making them highly susceptible to bias.Understanding the differences among these sampling methods helps statisticians choose the most appropriate technique for a study and recognize when a sample may be biased or unrepresentative.

The PGCC Math Department wants to survey students about use of its tutoring center. Since students are already divided into classes, the researchers randomly select 10 classes and give the survey to all the students in those classes.

You Try It 1.4.A

Identify the sampling technique used in this scenario:

Every 15th person entering the stadium was selected for an additional security screening.

Answer available at the end of this section.

You Try It 1.4.B

Identify the sampling technique used in this scenario:

A sample of faculty is selected so that it has equal representation: 10 assistant professors, 10 associate professors, and and 10 professors.

Answer available at the end of this section.

You Try It 1.4.C

Identify the sampling technique used in this scenario:

Viewers of a new TV show are asked to vote about how they liked the season finale on the show's website.

Answer available at the end of this section.

You Try It 1.4.D

Identify the sampling technique used in this scenario:

> An online retailer uses a computer to generate 50 account numbers and sends a satisfaction survey to the 50 customers associated with those account numbers.

Answer available at the end of this section.

You Try It 1.4.E

Identify the sampling technique used in this scenario:

> To survey students about how much homework they do on average, the PTA president randomly selects 10 school buses. Then, everyone who rides on those buses is interviewed.

Answer available at the end of this section.

You Try It 1.4.F

Identify the sampling technique used in this scenario:

> A pollster stands on a busy street corner and interviews the first 20 people who agree to talk with her.

Answer available at the end of this section.

Experiments

Samples and surveys called **observational studies** because they involve collecting and analyzing data without changing conditions or the individuals from which the data are collected. Data are also commonly collected as a result of an experiment. An **experiment** is a type of study in which the researcher deliberately changes a variable or imposes a treatment to understand a cause-and-effect relationship. It is common to use experiments when exploring how subjects react to an outside influence.

In an **experimental study**, variables of interest are identified. One (or more) of these variables is referred to as the **factor** of the study. The factor is manipulated by the researcher so that data may be obtained about its effect on another variable, called the **response variable**.

Consider an experiment to determine the effect of three different exercise programs on the cholesterol level of patients with high cholesterol. Each patient is referred to as an experimental unit, the response variable is the cholesterol level of the patient at the completion of the program, and the exercise program is the factor whose effect on cholesterol level is being investigated. The exercise program that each patient is assigned is referred to as a treatment.

Example 1.4.5

A pharmaceutical company tests a new medicine for Alzheimer's disease by administering the drug to 50 elderly patients with recent diagnoses and a placebo to 50 elderly patients with recent diagnoses.

- The factor is the new Alzheimer's medicine.
- The treatments are the new medicine and the placebo.
- The response variable is the effect on the patients.

You Try It $1.4.G$

Identify the response variable, the factor, and the treatments in this experiment:

A psychology researcher explores the effect of music on temperament by measuring people's mood while listening to rap, classical, or pop music.

Answer available at the end of this section.

Section Summary

Samples and surveys are forms of observational studies, and poor techniques in designing them can lead to incorrect conclusions about a population. When a sampling method produces a sample that does not accurately reflect the population, the sample is considered biased.

Three major sources of bias in samples are selection bias, response bias, and non-response bias. Selection bias occurs when the sample chosen fails to represent the population. Response bias arises when the survey itself influences how individuals answer. Non-response bias appears when many selected individuals do not participate, leaving the final sample unrepresentative.

Randomization is the key tool for minimizing bias. A simple random sample (SRS) ensures that each sample of size n has an equal chance of being selected, and technology can be used to simulate this random selection. Other sampling methods that incorporate randomization include systematic sampling, cluster sampling, and stratified sampling.

In contrast, convenience sampling and voluntary response sampling should be avoided because they do not use randomization and typically produce biased, unrepresentative samples.

Data may also be collected through an experiment, which attempts to determine whether explanatory variables, called factors, cause changes in a response variable. Experiments allow researchers to study cause-and-effect relationships that observational studies cannot establish.

Section 1.4 Review Questions

1. Brandy wants to know which brand of shoe high school soccer players prefer. She decides to ask the girls on her team which brand they like most.
   a. In Brandy's sample, which group(s) from the population is/are going to be under-represented?
   b. What type of bias is best illustrated in her sampling technique? Why?
   c. Brandy got a list of all the soccer players in the Colonial conference from her athletic director, Mr. Sprain. What is this list called?
   d. If she groups the list by boys and girls, and chose 40 boys at random and 40 girls at random, what type of sampling best describes her method?

2. Your doorbell rings, and you open the door to find a 6-foot-tall boa constrictor wearing a trench coat and holding a pen and a clip board. He says to you, "I am conducting a survey for a local clothing store. Do you own any boots, purses, or other items made from snakeskin?" After recovering from the initial shock of a talking snake being at the door, you quickly and nervously answer, "Of course not," as the wallet you bought on vacation last summer at Reptile World weighs heavily in your pocket. What type of bias best describes this ridiculous situation? Explain why.

3. In each situation, identify the type of bias (*selection bias, response bias*, or *non-response bias.*) Suggest why you think it is biased.
   a. A survey asks people to report their annual income for last year.
   b. A survey randomly calls people and asks them to answer a set of 75 questions that will take over an hour to complete
   c. A survey asks the following question: *Should the mall prohibit loud and annoying rock music in clothing stores catering to teenagers*
   d. To determine voter opinions about a downtown renovation project, a surveyor randomly questions people who work

downtown

e. A researcher wants to know whether students at a university have ever cheated on a test. The researcher walks up to 100 random students and asks them the following question: "*Have you ever cheated on a test or exam*?"

4. Researchers want to collect cholesterol levels of U.S. patients who had a heart attack. The following are different sampling techniques that the researcher could use. Classify each as *simple random sample*, *stratified sample, systematic sample, cluster sample, voluntary response,* or *convenience sample*.
   a. The researchers randomly select 5 hospitals in the U.S. then measure the cholesterol levels of all the heart attack patients in each of those hospitals.
   b. The researchers list all of the heart attack patients and measure the cholesterol level of every 25th person on the list.
   c. The researchers go to one hospital on a given day and measure the cholesterol level of the heart attack patients at that time.
   d. The researchers write the names of all of the heart attack patients on cards. They then mixed the cards in a box, pull out 40 cards, and measure the cholesterol levels of those selected patients.
   e. The researchers divide the heart attack patients based on race and measure the cholesterol levels of 10 randomly selected patients in each race grouping.

5. The quality control officer at a manufacturing plant needs to determine the average number of defective items per batch during the last month of production. The following are different sampling techniques that could be used by the officer. Classify each as *simple random sample*, *stratified sample, systematic sample, cluster sample, voluntary response,* or *convenience sample*.
   a. The officer lists all of the batches made last month and chooses some of those batches at random to count.
   b. The officer takes the first 10 batches and counts the defective items in each batch.
   c. The officer groups the batches made last month by which shift they were made. The number of defective items is counted in randomly selected batches in each shift.
   d. The officer chooses every 15th batch off the line and counts the number of defective items in each chosen batch.
   e. The officer divides the batches made last month by the day they were made. Then certain days are picked and every batch made on those days is counted to determine the number of defective items.

6. An amusement park wants to know if its new ride, *The Pukeinator*, is too scary. For each sampling scenario described below, i) identify the sampling method most evident, and ii) identify any bias you perceive in the choice of data collection.
   a. The first 30 riders of the day are asked their opinions of the ride.
   b. The name of a color is selected at random, and all people at the amusement park wearing that color are asked their opinion of the ride.
   c. A flier is passed out inviting people at the park to complete a survey about the ride when they return home.
   d. All tickets from the park are collected in a large box and 30 tickets are selected from the box after the park closes. After matching the tickets to phone numbers, calls are made to ask the 30 visitors some questions about the ride
   e. Every 12th person exiting the ride is asked in front of his friends: "*You didn't think that ride was scary, did you*?"
   f. Five riders are selected at random during each hour of the day, from 9 AM until closing at 5 PM.

7. There are 35 students taking physical science, and you want to sample some of them for a survey about their impressions of the course. You will use a 1-in-5 systematic sample. Assume the students are assigned numbers from 1 to 35. You choose the 2nd student as the first student in the sample. Which students are in the sample?

Answers

You Try It $1.4.A$

Systematic sample. A pattern is used to identify the sample from the population. We refer to this as a 1-in-15 systematic sample.

You Try It $1.4.B$

Stratified sample. The sample is being collected according to three strata based on faculty rank. Separate samples of assistant professors, associate professors, and professors were collected. The three smaller samples were then combined to form the stratified sample.

You Try It $1.4.C$

Voluntary response sample. The individuals in the sample are self-selected. A viewer initiates being included in the sample. This most likely results in bias.

You Try It $1.4.D$

Simple random sample. The account numbers are randomly generated. We have no reason to think that some customers have a better chance of being selected than others.

You Try It $1.4.E$

Cluster sample. The population has been divided into buses, which serve as geographic clusters. Once a bus is selected, everyone who rides on that bus is part of

the sample.

You Try It $1.4.F$

Convenience sample. The sample is collected from an area where it easy to reach people. If the people who pass by this corner are somehow different than their representation in the target population, we expect bias in the result.

You Try It $1.4.G$

- Response variable: people's temperament
- Factor: music
- Treatments: rap music, classical music, and pop music

This page titled 1.4: Collecting Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

2: SUMMARIZING DATA NUMERICALLY

2.1: Measures of Center
2.2: Measures of Position
2.3: Measures of Spread
2.4: Changes to a Data Set

This page titled 2: Summarizing Data Numerically is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

2.1: Measures of Center

Learning Objectives

- Calculate measures of central tendency.
- Explain conceptual differences among the mode, median, and mean.
- Apply the leveling-off and balance-point interpretations of the mean.
- Use notation for sample and population means.
- Describe the influence of outliers.
- Solve problems involving the mode, mean, and median.
- Calculate a weighted mean.

Introduction

Once data are collected, it is useful to summarize the data set by identifying a value around which the data are centered or clustered. This ‘center value’ is a value that we consider to be a *typical value,* one that is representative of the entire data set. Many times, we refer to this idea as an *average*, but we must be careful to specify what we are focusing on when calling something 'average.'

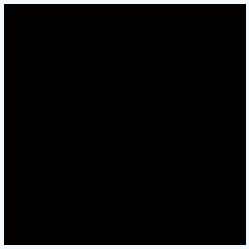
Three commonly used measures of center, or averages, are the mode, the median, and the mean. This section examines each measure of center based on the primary characteristic it uses to determine "averageness."

The Mode

The **mode** is a measure of center that identifies the value or category that occurs most frequently in a data set. The mode can be used with quantitative data or categorical data. Because it highlights the most common observation, the mode is especially useful for describing typical outcomes in situations where frequency is more meaningful than numerical magnitude, such as identifying the most popular class time, the most common shoe size sold, or the most frequently chosen survey response.

Example 2.1.1

Antoinette collected color data on the vehicles in the parking lot as show in the table. What is the mode of these data?



Solution

'Green' is the mode because it is the data value that occurred most often.

Example 2.1.2

The students in a statistics class were asked to report the number of children that live in their house (including brothers and sisters temporarily away at college). The data are recorded below. What is the mode?

1 3 4 3 1 2 2 2 1 2 2 3 4 5 1 2 3 2 1 2 3 6

Solution

The mode is 2 as it is the most frequently occurring number of children in the sample, telling us that more students in the class come from families where there are 2 children than from any other household size.

There may be one mode, more than one mode, or no mode at all. In the previous example, if there were seven 3-children households and seven 2-children households, we would say the data set has two modes, or the data set is *bimodal*. Technically, there can be more than two modes, but the more modes there are, the less useful the mode becomes. In these cases, we would most likely search for a different statistic to describe the center of such data. If each data value occurs only once or if there is an equal number of each data value, the mode is not useful in helping us understand the data, and thus, we say the data set has no mode.

You Try It 2.1.A

A small café recorded the number of smoothies sold each day for ten days. The data are shown below:

12 15 18 15 20 22 18 17 19 15

Find the mode number of smoothies sold.

Answer available at the end of this section.

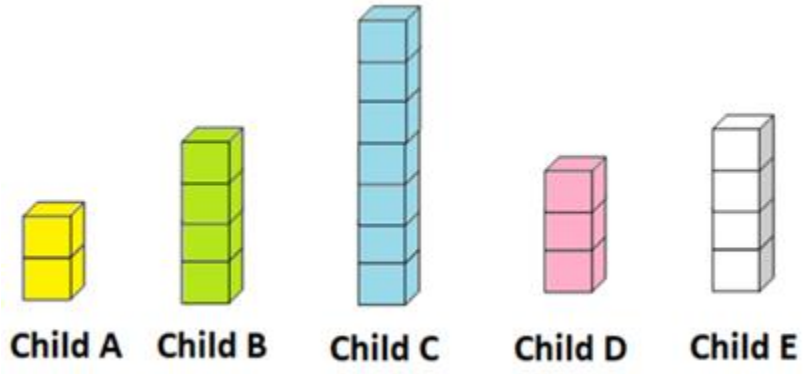
The Mean

Another measure of central tendency is the *arithmetic average* or mean. We will look at two interpretations of the mean that are often seen in elementary school curricula: *leveling out* and *balance point*.

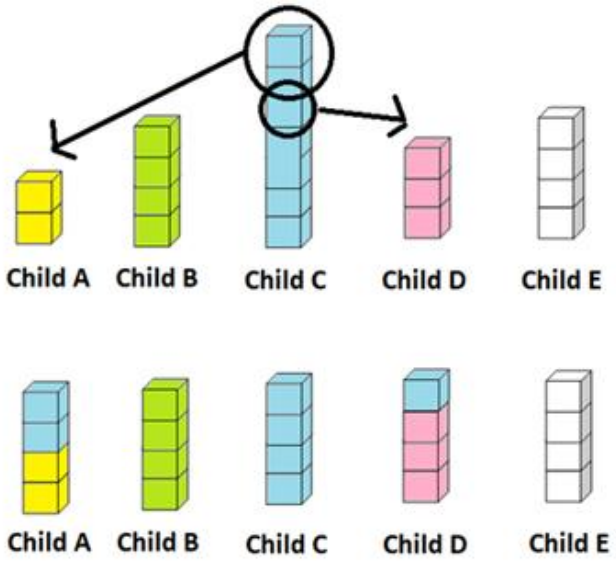
LevelingOut Interpretation of the Mean

One interpretation of the mean is *leveling out*. The mean is the value that results from equal sharing of a set of values among n individuals.

Suppose five children build towers of blocks. Their towers contain 2, 4, 7, 3, and 4 blocks.



If the children redistribute the blocks so that each child ends up with the same number, blocks would be moved from the taller towers to the shorter ones.



After leveling, each child would have 4 blocks. This shared value, 4, is the mean.

Notice that the total number of blocks (20) and the number of children (5) never change—only the distribution does. The original data set {2, 4, 7, 3, 4} becomes {4, 4, 4, 4, 4} once the values are leveled. Because each child ends with 4 blocks, the mean is 4.

Another way to imagine equal sharing is to place all 20 blocks in a single pile and then distribute them evenly until none remain. This mirrors the familiar procedure for computing the mean: add all the data values to find the sum, then divide by the number of values. Symbolically, the formula for the sample mean is

$$\bar{x} = \frac{\dsum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{n}$$

where x_i is the ith data value and n is the sample size. Statisticians use the symbol $\bar{x}$ to represent the mean while x is the symbol for a single measurement. Say $\bar{x}$ as "*x bar*."

The formula for finding the mean of a population is the same, but we denote that it was calculated through a census by the Greek letter, μ. Say μ as "*mew*."

$$\mu = \frac{\dsum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_N}{N}$$

The sample mean $\bar{x}$ is a statistic because it is measuring a sample, and μ is a parameter because it is measuring a population. We say that $\bar{x}$ is a point estimate of μ.Balance Point Interpretation of the Mean

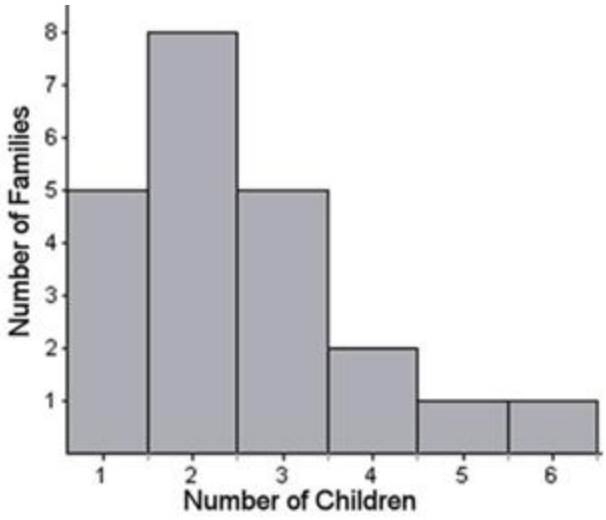
Balancing Point Interpretation of the Mean

Another common interpretation of the mean is as the balancing point of a distribution. We can illustrate this geometric interpretation of the mean as the balance point. Recall the number of children in a household data from Example 2.1.2:

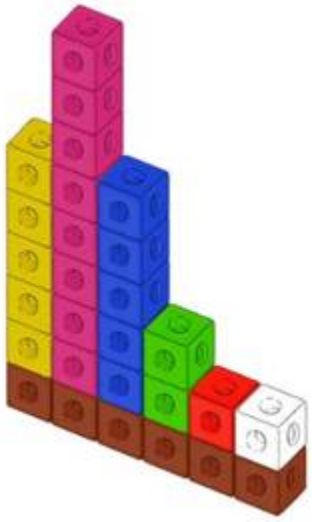
1 3 4 3 1 2 2 2 1 2 2 3 4 5 1 2 3 2 1 2 3 6

There are $n = 22$ students in this class, and the total number of children in all their houses is $\sum x = 55$, so the mean of this data is $\bar{x} = \frac{55}{22} = 2.5$ children.

Here is a graph of that data:

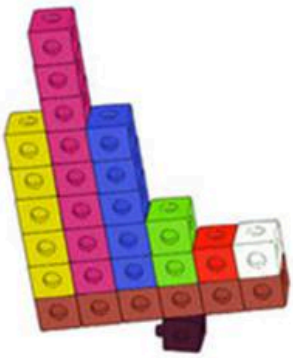


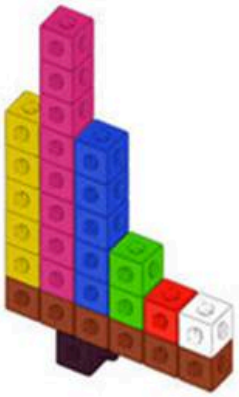
We can physically model the graph with snap cubes using one cube to represent each student's family and a row of six cubes at the bottom to hold them together:



It turns out that the model of blocks *balance*s at 2.5. In the pictures below, you can see that a block placed at 2 makes the graph to tip to the right, while one placed at 3 causes the graph to tip to the left. However, if you place the block at 2.5, it balances perfectly!







You Try It $2.1.B$

A small café recorded the number of smoothies sold each day for ten days. The data are shown below:

12 15 18 15 20 22 18 17 19 15

Find the mean number of smoothies sold.

Answer available at the end of this section.

You Try It $2.1.C$

Six friends gathered small river stones during a nature walk. Once they finished their walk, they decided to redistribute the stones so that each friend has 5 stones. Based on this information:

a. What is the mean number of stones the friends collected?
b. How many stones did the friends collect in all?

Answer available at the end of this section.

The Median

The **median** is simply the number in the middle position in an *ordered* set of data.

Suppose a student took five statistics quizzes and received the following scores:

80 94 75 96 90.

To find the median, first put the data in order. The median will be the score that is in the middle. Placing the data in order from least to greatest yields

75 80 90 94 96.

The number in the middle position is the third score, or 90, so the median of this data is 90.

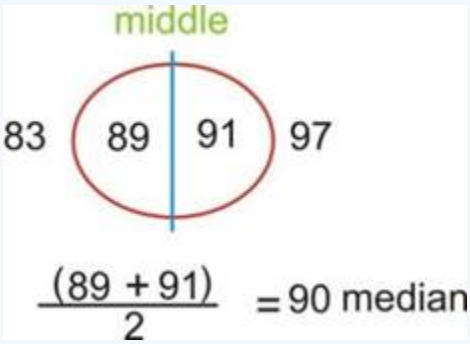
However, when there is an even number of data values, there won't be *just one* data value in the middle. In that case, we find the mean of the two data values that fall in the middle positions.

Example 2.1.3

Find the median of the following quiz scores: 91, 83, 97, 89.

Solution

Once the numbers are ordered numerically, the numbers in the second and third positions straddle the middle of this set. The mean of these two numbers is 90, so the median of the data is 90.



While it is easy to find the "middle" of a small set of numbers just by looking, this is a bit more trouble when the data set is large. It may be useful to have a guide to help.

Finding the Median of a Large Data Set

1. List the data in order from smallest to largest, or largest to smallest.
2. If the number of data values n is *odd*, then the median will be the data value in the middle position. To find the position of the median when n is odd, round $n/2$ up to the next whole number and count until you reach that position.
3. If the number of data values n is *even*, there are two values in the middle positions. Find the mean of the values in positions #$(\frac{n}{2})$ and #$(\frac{n}{2}+1)$.

Example 2.1.4

Let's revisit the sample data about number of children in the household from Example 2.1.1. Find the median.

1 3 4 3 1 2 2 2 1 2 2 3 4 5 1 2 3 2 1 2 3 6

Solution

Begin by ordering the data.

1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 4 4 5 6

There are $n = 22$ data values, which is even. Find the mean of the values in positions #$\left(\frac{22}{2} = 11\right)$ and #$\left(\frac{22}{2} + 1 = 12\right)$. These two values are 2 and 2. so the median is $\frac{2+2}{2} = 2$.

You Try It 2.1.D

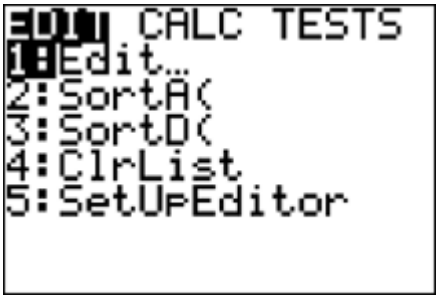
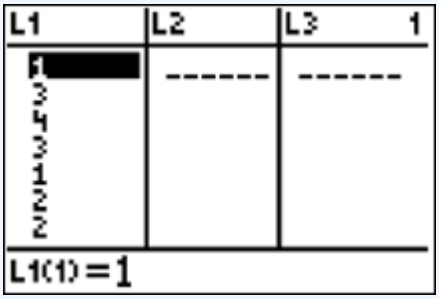
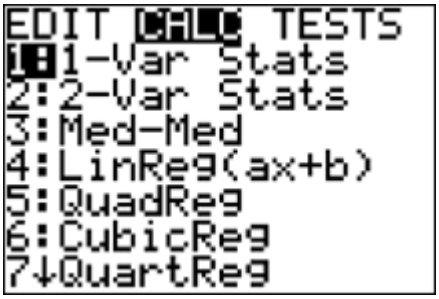
A small café recorded the number of smoothies sold each day for ten days. The data are shown below:

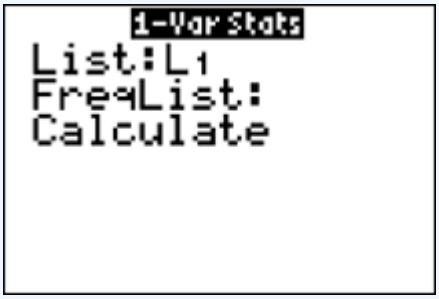
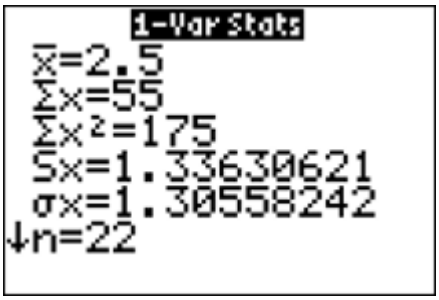
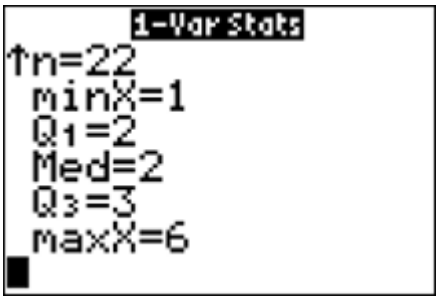
12 15 18 15 20 22 18 17 19 15

Answer available at the end of this section.

For larger data sets, technology is useful. The procedure is shown in the table below.

Finding the Mean and Median on the TI-83/84 Calculator

1.	Press **[STAT]** to open the Statistics menu. Choose **1:Edit** on the **EDIT** menu. Press **[ENTER]**.	
2.	You now have a screen that allows you to enter the data in a column. Enter the number of children data in L1. Be sure to press **[ENTER]** after each value.	
3.	Again, press **[STAT]** and use the right arrow to access the **CALC** menu.	

4.	Choose **1:1-Var Stats** on the CALC menu. You will have one of the two screens shown: • If you have a screen like the one on the top, press **[2ND] [1]** for the list L1 and press **[ENTER]**. • If you have a STAT WIZARD screen that looks like the one on the bottom, enter the name of the list by pressing **[2ND] [1]**. Leave **FreqList** blank. Highlight **Calculate** and press **[ENTER]**.	 
5.	The calculator will display several values. We will learn about all these statistics later. The first screen shows the mean $\bar{x}$, the sum of the data values $\sum x$, and the number of data values n. If you use the down arrow, you will also see the median (Med).	 

Example 2.1.5

This table shows the number of home runs for the past ten baseball games of the season. Find the median number of home runs.



Solution

One strategy would be to list all the data and proceed as we did earlier. However, it is more efficient to work in the table since the data are already ordered.

There are $n = 1 + 2 + 0 + 2 + 4 + 1 = 10$ games so $n = 10$, which is even. Find the data values in the $\left(\frac{10}{2} = 5\right)^{\text{th}}$ position and the $\left(\frac{10}{2} + 1 = 6\right)^{\text{th}}$ position. Count to find the data values in these positions and average them. There are 7 runs in the 5th position and 8 runs in the 6th position. So, the median is $\frac{7+8}{2} = 7.5$ home runs.

You Try It $2.1.E$

The table shows how many workout sessions members of a gym completed last month. Find the mode and median number of workout sessions.



Answer available at the end of this section.

The Mean vs. the Median

Both the mean and the median are important and widely used measures of center for quantitative data. Is one better to use than the other? It depends -- as the following example shows.

Example $2.1.6$

Suppose you earned an 85 and a 93 on your first two statistics exams, but then you had a really bad day and got a 14 on your next exam!

The mean of your three scores is 64: $\frac{85 + 93 + 14}{3}$.

The median of your three scores is 85: 14 **85** 93

Which do you think gives a more representative picture of your performance?

Solution

The median does not change if the lowest grade is an 84 or if the lowest grade is a 14. However, when you add the three numbers to find the mean, the sum will be much smaller if the lowest grade is a 14. The mean is significantly lower than the median because of that one low score. The more representative statistic is a matter of opinion, but this exam does show that the mean and median can behave differently due to the one bad day performance.

The mean and the median can behave very differently when a data set contains a value that is far from the rest. In statistics, such an unusually large or small value is called an **outlier**. Outliers can strongly influence some measures of center while leaving others largely unchanged.

The mean is sensitive to outliers because it incorporates every value in the data set through addition. A single extreme value can pull the mean toward it, making the mean **non-resistant**. In contrast, the median depends only on the position of values when the data are ordered, not on their magnitude. Because an outlier does not change the middle position, the median is typically **resistant** to extreme values.

A **resistant statistic** is one that is not significantly affected by outliers. A **non-resistant statistic** is one that is strongly influenced by them. The median is a resistant measure of center, while the mean is non-resistant. In a sense, the median can "hold its ground" against the pull of an extreme value, but the mean is drawn toward that value because it must account for it in the calculation.

When a data set contains outliers, it is often more appropriate to use the median to describe the center. The median provides a more stable and representative summary of the typical value in situations where extreme observations distort the mean.

Example 2.1.7

A large tech company reports employee salary data for the year. Most software engineers, designers, and support staff earn between $70,000 and $150,000. However, the company's CEO earns over $50 million due to stock awards and bonuses. Which measure of center - the mean or the median - best represents a typical salary at this company?

Solution

This single, extremely high salary is not representative of what typical employees earn. Because the CEO's compensation is so far from the rest of the data, it dramatically increases the mean salary and makes it appear as though employees earn far more than they actually do. In this situation, the median salary provides a much more accurate picture of what a typical employee earns, since the median is resistant to the influence of such an extreme outlier.

There are many circumstances when one would prefer to use the median over the mean. Here are some real-life situations where the median is the better choice:

- Income or salary data A few very high earners (executives, celebrities, founders) can pull the mean upward, making it look like people earn more than they do. The median salary shows what a typical worker earns.
- Test scores with one very low or very high score — If one student leaves a test blank or one student scores perfectly, the mean shifts. The median remains stable.
- Home values in a neighborhood — One mansion among modest homes will inflate the mean. The median shows the typical home value.

Problem Solving with the Mean

Typically, we think about finding the mean from a set of data that is given or collected. But there may also be situations where you need the mean to be a certain value and want to find a missing data value or even create the entire data set. We need to think about the formula for the mean flexibly.

We can rewrite the formula for the mean algebraically to solve for the sum of the data values:

$$\bar{x} = \frac{\sum x}{n}$$
$$n \cdot \bar{x} = \sum x$$
$$\sum x = n \cdot \bar{x}$$

This shows that the sum of all the data values $sumx$ equals the product of the number of data values and the mean,(\n \cdot \bar{x}\). This is not surprising. Recall that in the leveling out interpretation of the mean, the total number of blocks always stays the same and will be equal to the number of stacks (n) times the number of blocks in each stack after they are leveled $(\bar{x})$.

Example 2.1.8

The 28 students in Ms. Tracy's homeroom collected canned food items for the food drive. If the mean number of cans collected by each student was 12, how many cans of food did the class collect?

Solution

While we don't know how many cans any the individual 28 students collected, we know that the mean was 12 cans. This tells us that $n = 22$ and $\bar{x} = 12$. Substituting into the rearranged formula for the sum of the data gives

$$\begin{aligned} \sum x &= n \cdot \bar{x} \\ &= 28 \cdot 12 \\ &= 336 \end{aligned}$$

Using leveling out thinking, there are 28 stacks with 12 cubes in each stack after being leveled. So, the total of cubes in all the stacks before being leveled is the same as $28 \cdot 12 = 336$.

You Try It 2.1.F

The 35 volunteers at the Riverside Animal Shelter each walked several dogs last weekend. If the mean number of dogs walked per volunteer was 5, how many dogs were walked in total?

Answer available at the end of this section.

Sometimes we use a similar process to find a missing individual data value, x, if we know the mean.

Example 2.1.9

To qualify for the bowling tournament, Mara must have at least a 180 average bowling score in her most recent 5 games. So far, Mara has bowled 4 games with scores of 170, 184, 160, and 195. What score must Mara get in her 5th game to have a mean of 180?

Solution

Again, we know $n = 5$ and $\bar{x} = 180$. We also know four of the individual data values but not the fifth one: $x_1 = 170$, $x_2 = 184$, $x_3 = 160$, $x_4 = 195$, and $x_5 = ?$.

Substituting into the rearranged formula for the sum of the data gives

$$\sum x = n \cdot \bar{x}$$
$$170 + 184 + 160 + 195 + x_5 = 5 \cdot 180$$
$$709 + x_5 = 900$$
$$x_5 = 191$$

Mara needs to score 191 to bring her bowling average to 180.

You Try It 2.1.G

To qualify for the city cycling time-trial, Luis must have at least a 24-minute average in his most recent 6 races. So far, Luis has completed 5 races with times of 26, 22, 25, 23, and 27 minutes.

What time must Luis get in his 6th race to have a mean of 24 minutes?

Answer available at the end of this section.

Example 2.1.10

Design a set of data to represent nine 10-point quiz scores where the mean is 7 and the median is 8.

Solution

The sum of the nine quiz scores must be $9 \times 7 = 63$ points. There are an odd number of data values and the median is 8 so the middle value must be 8.

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad 8 \quad x_6 \quad x_7 \quad x_8 \quad x_9$$

The missing data values must all add to be $63 - 8 = 55$, the x-values to the left of 8 must be 8 or less and the x-values to the right of 8 must be 8 or more (but 10 or less.) There are many possibilities. Here is one possible solution: 2, 5, 6, 7, 8, 8, 8, 9, 10.

Example 2.1.11

Mr. Henry gave the same science test to two classes. The morning class of 18 students scored a mean of 78 points. His afternoon class of 24 students scored a mean of 85 points. What is the mean score when the classes are combined into one group?

Solution

First, calculate the sum of the scores for each class using the rearranged mean formula:

$$\dsum x_{\text{morning}} = n \cdot \bar{x} = 18 \cdot 78 = 1440$$

$$\dsum x_{\text{afternoon}} = n \cdot \bar{x} = 24 \cdot 85 = 2040$$

Next, find the mean of the combined scores by adding the sums for the two classes into one sum and divide by the total number of students in the two classes together:

$$\bar{x}_{\text{combined}} = \frac{\sum x_{\text{morning}} + \sum x_{\text{afternoon}}}{n_{\text{morning}} + n_{\text{morning}}} = \frac{1404 + 2040}{18 + 24} = \frac{3444}{42} = 82$$

When considered as one group, the mean score of the two classes is 82.

The Weighted Mean

The previous example illustrates a **weighted mean**. We cannot find the overall mean for both classes by simply "averaging the two averages" because the afternoon class has more students than the morning class. Its average must therefore carry more influence in the combined result. To obtain an accurate mean for all students together, we must give the larger class greater weight in the calculation, ensuring that each individual's score contributes appropriately to the final average.

One common type of weight to use is the frequency, which is the number of times each number is observed in the data. When we calculated the mean for the children in a home in Example 2.1.2, we could have used a weighted mean calculation. The calculation would look like this:

$$\bar{x} = \frac{(5 \cdot 1) + (8 \cdot 2) + (5 \cdot 3) + (2 \cdot 4) + (1 \cdot 5) + (1 \cdot 6)}{5 + 8 + 5 + 2 + 1 + 1} = \frac{55}{22} = 2.5$$

A symbolic representation of the weighted mean is

$$\bar{x}_{\text{weighted}} = \frac{\sum f_i \cdot x_i}{\sum f_i}$$

where x_i is the ith data value and f_i is the number of times that data value occurs.

Example 2.1.12

The results on a 20-point multiple choice test are as follows: nine students scored 20 points, four students scored 19, three students scored 18, five students scored 17, and one student scored 12. What was the mean score for the class?

Solution

The scores were 20, 19, 18, 17, and 12. However, we cannot just average those scores because they occurred with different frequencies. The scores must be weighted by the number of times they each occurred. Using the formula for the weighted mean,

$$\begin{aligned}
\bar{x}_{\text{weighted}} &= \frac{\sum f_i \cdot x_i}{\sum f_i} \\
&= \frac{(9 \cdot 20) + (4 \cdot 19) + (3 \cdot 18) + (5 \cdot 17) + (1 \cdot 12)}{9 + 4 + 3 + 5 + 1} \\
&= \frac{180 + 76 + 54 + 85 + 12}{22} \\
&= \frac{407}{22} \\
&= 18.5
\end{aligned}$$

The class mean is 18.5 points.

You Try It 2.1.H

Two local running clubs tracked how many miles their members ran last week.

- The River Runners Club, with 16 members, ran an average of 12 miles each.
- The Hilltop Striders Club, with 20 members, ran an average of 15 miles each.

What is the mean number of miles run when the two clubs are combined as one group?

Answer available at the end of this section.

You can also weight data by a proportion (such as a fraction, decimal, or percentage.) You are sure to have experienced this many times in calculating your own grades. Here, the frequency f_i is replaced by p_i in the formula.

Example 2.1.13

Professor Henry uses a weighted grade system to determine his students' grades in a history class. Tests count 50% of the course grade, while quizzes count 15%, a research paper counts 10%, and the final exam counts 25%.

Jessica has an 82% average on tests, 90% average on quizzes, 78% on her research paper, and a 72% on the final exam. What is Jessica's course average?

Solution

We will use the formula for a weighted mean where percentages p_i are used as the weights instead of frequencies f_i.

$$\begin{aligned}\bar{x}_{\text{weighted}} &= \frac{\sum p_i \cdot x_i}{\sum p_i} \\ &= \frac{(50\% \cdot 82\%) + (15\% \cdot 90\%) + (10\% \cdot 78\%) + (25\% \cdot 72\%)}{50\% + 15\% + 10\% + 25\%} \\ &= \frac{(0.50 \cdot 0.82) + (0.15 \cdot 0.90) + (0.10 \cdot 0.78) + (0.25 \cdot 0.72)}{0.50 + 0.15 + 0.10 + 0.25} \\ &= \frac{0.41 + 0.135 + 0.078 + 0.18}{1} \\ &= \frac{0.803}{1} \\ &= 0.803\end{aligned}$$

Jessica's course average is 80.3%.

You Try It 2.1.*I*

A music academy uses a weighted system to determine each student's final performance score:

- Instrument Technique counts 55% of the final score.
- Music Theory counts 30%.
- Ensemble Performance counts 15%.

Evan earned the following scores in each category: Instrument Technique 91%, Music Theory 84%, and Ensemble Performance 88%.

What is Evan's final performance score?

Answer available at the end of this section.

Another common use of weighted mean is calculation of a student's grade point average (GPA.) The GPA is a numerical summary of academic performance. Each course grade is converted into a number on a standard scale—typically 0.0 to 4.0—and these values are combined using a weighted mean, giving more weight to courses with more credit hours. At the college level, students earn points for each grade they receive: A = 4 points, B = 3 points, C = 2 points, D = 1 point, and F = 0 points.

Example 2.1.14

Here is Tiffany's grade report for the Fall semester. Find Tiffany's GPA.



Solution

Use the weighted mean formula where f_i is the number of credits each course is worth and x_i is the number of points Tiffany earned for the grade.

$$\begin{aligned}\bar{x}_{\text{weighted}} &= \frac{\sum f_i \cdot x_i}{\sum f_i} \\ &= \frac{(5 \cdot 4) + (3 \cdot 1) + (4 \cdot 3) + (1 \cdot 2) + (1 \cdot 4)}{5 + 3 + 4 + 1 + 1} \\ &= \frac{20 + 3 + 12 + 2 + 4}{5 + 3 + 4 + 1 + 1} \\ &= \frac{41}{14} \\ &\approx 2.93\end{aligned}$$

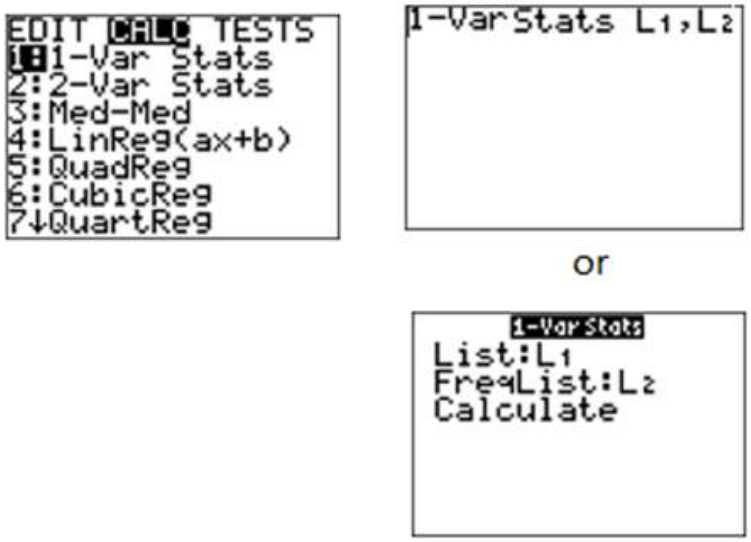
Tiffany's GPA for the Fall semester was 2.93.

Technology Note: Finding a Weighted Mean on the TI-83/84 Calculator

If you have weights for data values instead of the raw data when finding the mean and median, you can enter the frequency or weights in L2.

When following the previously shown steps for finding the mean and median, type L1,L2 following the 1:1-Var Stats command in the CALC menu.

If you have a STATS WIZARD screen, enter L2 for the FreqList.



Section Summary

When examining a set of data, descriptive statistics help us understand where the data are centered. The **mode**, **mean**, and **median** are the three most commonly used measures of center, each offering a different perspective on what is “typical” in a data set.

- Mode: The mode identifies the most frequently occurring value(s) in a data set. It is especially useful for categorical data or data measured at the nominal level, where numerical averaging is not meaningful.
- Mean: The sample mean, $\bar{x}$, is found by adding all data values and dividing by the number of values. It can be interpreted as the “leveling-out” amount or the physical balancing point of the data.
- Median: The median is the value that falls at the middle position of an ordered data set. With an odd number of observations, it is the middle data value; with an even number, it is the mean of the two middle values.

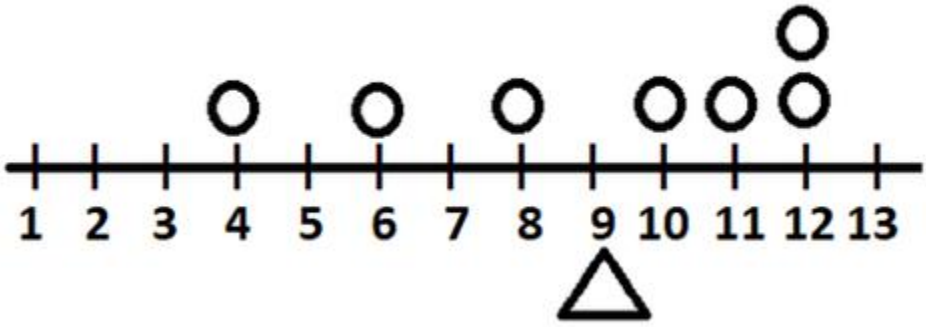
An outlier is an observation that is unusually large or small compared with the rest of the data. Outliers can strongly influence some measures of center. The median is a resistant statistic -- it is not significantly affected by outliers. The mean is non-resistant, meaning extreme values can pull it away from the typical center. For data sets containing outliers, the median often provides a more appropriate description of center.

A weighted mean is used when different data values contribute unequally to the final average. Weighted means are common when combining data sets of unequal size, summarizing grades, or calculating GPA, where values with greater frequency or credit hours must carry more influence in the overall result.

Section 2.1 Review Questions

1. Draw a sketch of blocks and explain how you can use the idea of *leveling out* to find the mean of each set of numbers below:
   a. {3, 6, 4, 7}
   b. {2, 1, 6, 5, 1, 6}

2. Jerome shows the data set {4, 6, 8, 10, 11, 12, 12} by placing counters on a number line. Jerome knows that the mean is 9 and shows the *balance point* is there.



   a. Where can Jerome put *one more* counter so that the balance point will not change?
   b. Where can Jerome put *two more* counters so that the balance point will not change?
   c. Suppose that Jerome puts one counter at 5 and another counter at 12. Where can Jerome put a third counter so that the balance point will not change?
   d. Suppose Jerome puts a counter at 2. Where could Jerome place two more counters so that the balance point will not change?

3. The following data represent the number of pop-up advertisements a user received while surfing the internet during the past month:

43 37 35 30 41 23 33 31 16 21

   a. Calculate the mean and median number of these advertisements by hand.
   b. Confirm the answers to part (a) using technology.

4. Find the mean, median, and mode of the following numbers. Which of them do you think gives the best 'average'? Why?

15 19 15 16 11 11 18 21 9 165

11 20 16 8 17 10 12 11 16 14

5. The table shows the heights of a group of 8th graders. Find the mode, mean, and median.



a. Calculate the mode, mean, and median by hand.
b. Confirm the answers to part (a) using technology.

6. In Lois' 2nd grade class, all the students are between 45 and 52 inches tall, except one boy, Lucas, who is 62 inches tall. Which of the following statements do you expect to be true about the mean and median heights of the students?

   - The mean height and the median height are about the same.
   - The mean height is greater than the median height.
   - The mean height is less than the median height.

7. Enrique has earned grades of 91, 87, and 95 in statistics for the first three quarters. To be exempt from taking the final exam, his mean grade for the year must be at least 93. Assuming grades are rounded using standard mathematical procedures, determine the lowest whole-number grade Enrique can earn in the fourth quarter that will give him a yearly mean of 93 or higher.

8. A math student scored 75, 70, 85, 90, and 100 on the first five tests he took. After he took his sixth math test, the average is now 85. What did he score on the sixth test?

9. The average time a man spent watching TV daily for the past week is 4 hours. If we eliminate one of these days, the average time he spent watching TV becomes 3.5 hours. How many hours did the man watch TV on the day we removed?

10. Create a set of 7 test scores (0 to 100 points) that have a mean of 82 and a median of 80.

11. Clarice rolled a number cube (labeled with the numbers 1 through 6) ten times and found the mean was 3 and the median was 3.5. Give one possible list of numbers she could have rolled.

12. The ages of the 8 children on a playground have a mean of 8.25 and a median of 8. There is no mode. Give a reasonable set of data for this situation.

13. Six people on an elevator have a mean weight of 110 pounds. Three more people with a mean weight of 150 pounds got on the elevator when the elevator stopped. What is the mean weight of all the people on the elevator?

14. For the month of April, a checking account has a balance of $700 for 24 days, 1215 for 2 days, and 375 for 4 days. What is the average daily balance for the month of April?

15. Mary ordered dinner for a party of 10 people. Three people ordered the $4.75 chicken dinner, two people ordered the $4.95 fish dinner, and five had the beef dinner at a cost of $6.75 each. What was the average cost of each dinner at Mary's party?

16. In a town with three elementary schools, all the fourth graders took a standardized test, with the results as shown in the table. What is the mean test score of all 4th graders in this town?

Standardized Test Averages by School



17. Sharla pretty much knows what grades she is going to get at the end of the semester. She expects her report card to look like this:

Shayla's Expected Grade Report



    a. Find Sharla's semester GPA based on this information.
    b. As it turned out, Sharla surprised herself and earned a B in College Algebra. Unfortunately, she only earned a B in Biology Lab. Will her GPA be higher or lower than the GPA you computed in part a? Why must this be so?

18. Mrs. Morris weights her course grades. Tests are worth 50%, quizzes are worth 15%, homework is worth 10% and the final project is worth 25%.
    a. At the end of the semester, Mike has a test average of 90, a quiz average of 87, a homework average of 65, and a final project average of 70. What is Mike's course grade?
    b. Julia has not completed her final project yet, but knows her test average is 82, her quiz average is 75, and her homework average is 92. What is Julia's grade before handing in her final project?
    c. Referring to part b, what score must Julia earn on her final project to earn a course grade of 80 or higher?

Answers

You Try It $2.1.A$
15 smoothies

You Try It $2.1.B$
17.1 smoothies

You Try It $2.1.C$
a. 5
b. 30

You Try It $2.1.D$
17.5 smoothies

You Try It $2.1.E$
mode = 2; median = 3

You Try It $2.1.F$
175

You Try It $2.1.G$
21 minutes

You Try It $2.1.H$
≈ 13.67 miles

You Try It $2.1.I$
88.45%

This page titled 2.1: Measures of Center is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

2.2: Measures of Position

Learning Objectives

- Interpret percentile and percentile rank in context.
- Find a percentile within a data set.
- Determine quartiles of a data set.
- Present the five-number summary of a data set.
- Identify outliers using the interquartile range (IQR.)

Introduction

Measures of center—such as the mean, median, and mode—are important tools for summarizing data, but they do not tell the whole story. For example, suppose the mean score on a social studies exam is 80%. From this single number, we cannot determine how the scores are distributed. It is possible that most students scored close to 80%, perhaps between 78 and 82. It is equally possible that no one scored near 80%, with all scores clustered above 95 or below 65. The mean alone cannot distinguish between these very different situations.

To gain a clearer understanding of a data set, we must look beyond the center. Measures of position and measures of spread help describe how the data are distributed and where individual values fall relative to the whole. In this section, we focus on measures of position, and in the next section we will explore measures of spread to complete the picture of how data behave.

Percentile and Percentile Rank

Percentiles and percentile ranks arise whenever we want to describe how a single value compares to an entire group. Raw numbers alone often fail to show whether a score is unusually high, unusually low, or somewhere in the middle. For example, knowing that a student scored 82 on an exam does not reveal how that score stands relative to everyone else. A score of 82 might be excellent in one class and below average in another. Percentiles help by identifying specific cutoff values within an ordered dataset, showing where important dividing points occur. Percentile ranks complement this by expressing an individual's standing as a percentage of observations at or below that score. Together, percentiles and percentile ranks provide a clear, standardized way to interpret data, compare performances, and understand position within a distribution.

Percentiles

As we have seen, the median is the value in a data set such that half of the observations are at or below it and where half are at or above it. For this reason, the median is also called the 50th percentile, and it serves as a central reference point in the distribution.

A *percentile* generalizes this idea. For example, the 90th percentile of a numerical data set is the value such that about 90% of the observations are less than or equal to it and the remaining 10% are greater. Other percentiles can be specified as well. A **percentile** identifies a specific value in the ordered data set so that a certain percentage of the data lie at or below that value and the rest lie above it.

There is no single universal definition or formula for computing percentiles, so different textbooks or software packages may use slightly different procedures. In this book, the pth percentile is defined as the daya value such that $p\%$ of the observations are less than or equal to it and the remaining $(100-p)\%$ are greater.

For instance, consider a set of exam scores. The 30th percentile is the score at which 30% of students scored the same or lower. It separates the lowest 30% of scores from the highest 70%. It does not indicate whether the score is good or bad. The percentile simply identifies the score's position within the distribution. When asked for a percentile, you will report a value from the data. Percentiles are written using whole-number labels—such as 10th, 30th, or 75th—rather than percentages.

Percentiles are measures of position that help describe how individual data values compare to the rest of a distribution. They are especially useful when working with large data sets, where measures of center alone -- such as the mean or median -- do not provide enough detail about how values are spread out. Percentiles can be used when raw scores are difficult to interpret on their own, such as test scores, waiting times, or measurements taken on different scales.

Example 2.2.1

During a basketball season, a player's average score falls at the 40th percentile. Interpret the meaning of the 40th percentile in the context of this situation.

Solution

A player whose average score falls at the 40th percentile is being compared to all other players in the league or group. The 40th percentile means that 40% of players have average scores that are less than or equal to this player's average, while 60% of players have higher averages.

The player scores higher than many players, but more players (about 60%) have higher averages. It places the player slightly below the middle of the overall scoring distribution.

You Try It 2.2.A

A teacher records the reading-fluency scores of students in her class. One student's score is at the 85th percentile.

a. This means the student scored as high or higher than _____% of the other students.
b. It also means that _____% of the students scored higher than this student's score.

Answer available at the end of this section.

Percentile Rank

A percentile rank is closely related to a percentile but describes something different. A percentile is a specific score in an ordered data set -- the value at which a certain percentage of observations fall at or below it. In contrast, a *percentile rank* reports a percentage that indicates how an individual score compares to the rest of the distribution. A score's **percentile rank** tells us what proportion of the data is at or below that score.

To see the distinction, suppose a teacher orders the test scores of 100 students from lowest to highest. The 60th percentile of this dataset is 78 points, meaning that 78 is the score at or below which 60% of all students scored. This value (78 points) is a percentile. It is a specific cutoff point in the data.

Now consider a student named Maya, who scored 82 points on the exam. After comparing her score to the rest of the class, the teacher determines that Maya scored better than 75% of all other students. Maya's percentile rank is 75%, indicating that her score is higher than 75% of the scores. This percentile rank describes Maya's standing relative to the group.

Example 2.2.2

A manufacturing engineer measures the lifespans (in hours) of 80 lightbulbs produced in a factory and arranges the lifespans from shortest to longest. One particular bulb lasts 510 hours. After reviewing the ordered data, the engineer finds that 62 of the 80 bulbs have lifespans less than or equal to 510 hours.

Compute the bulb's percentile rank and interpret the percentile rank in a complete sentence.

Solution

There are 80 bulbs, and 62 have lifespans less than or equal to 510 hours. The perecentile rank is computed as

$$\frac{62}{80} = 0.775 = 77.5$$

The bulb's percentile rank is 77.5, which is typically rounded to 78. A lifespan of 510 hours is at or above 78% of all measured bulb lifespans. In other words, this bulb lasted longer than about four-fifths of the bulbs in the sample.

So, percentiles and percentile ranks offer complementary perspectives:

- A percentile is a location in the data and a specific score.
- A percentile rank is a percentage describing how an individual score compares to the group.

So far we have emphasized the terminology and interpretation of percentile and percentile rank. We now turn to an example that allows you to see how these concepts operate within an actual distribution.

Example 2.2.3

A wildlife researcher records the number of berries eaten by 12 chipmunks in one afternoon:

3 4 5 6 6 7 8 9 10 11 12 14

Find the 40th percentile of this dataset and determine the percentile rank of a chipmunk who eats 9 berries.

Solution

The data are already in ordered form, which is required before identifying percentiles.

To locate the 40th percentile, determine the position in the ordered list corresponding to 40% of the 12 observations:

$$0.40 \times 12 = 4.8.$$

This means the 40th percentile lies at approximately the 5th observation in the ordered data. The 5th value is 6, so the 40th percentile is 6 berries. About 40% of the chipmunks ate 6 or fewer berries, and the remaining 60% ate more than 6 berries.

To find the percentile rank for the chipmunk that eats 9 berries, count how many data values are less than or equal to 9. There are 8 such values out of 12 total observations. Converting this to a percentage shows that

$$\frac{8}{12} \approx 67\%.$$

Thus, a chipmunk who eats 9 berries has a percentile rank of approximately 67. A chipmunk that eats 9 berries eats more than about two-thirds of the group.

You Try It 2.2.B

A botanist measures the heights of 10 young sunflower plants in centimeters:

14 16 18 19 21 23 25 28 30 33

a. Find the 60th percentile of this dataset.
b. Determine the percentile rank of a sunflower that is 18 cm tall.

Answer available at the end of this section.

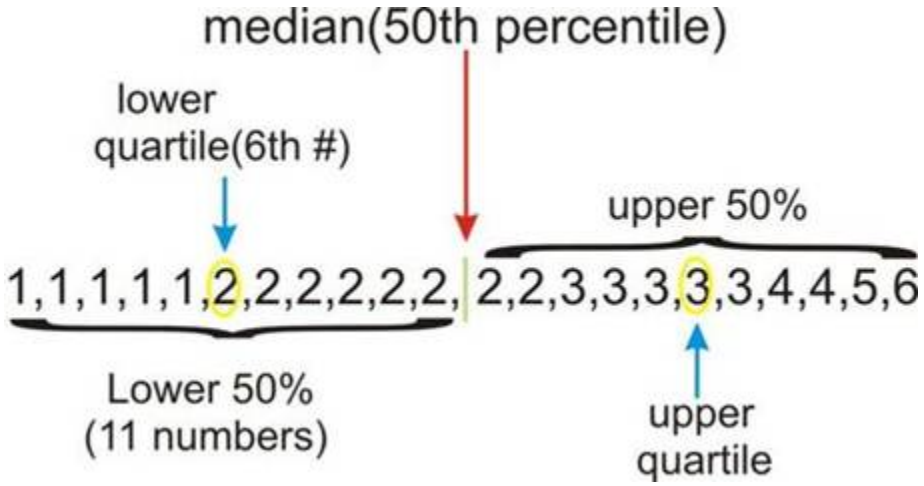
Quartiles

Three very common percentiles are the 25th, 50th, and 75th percentiles. The 25th percentile, 50th percentile (median), and 75th percentiles are points that divide the data into four quarters. Because of this, the 25th percentile is notated as Q_1 and is called Quartile 1. The 75th percentile is notated as Q_3 and is called Quartile 3. The median is the second quartile and is sometimes referred to as Q_2.

Let's return to the ordered data set about the number of children in a household from the last section:

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 6

Recall that the median (50th percentile) is 2. You can think of Q_1 and Q_3 as the medians of the upper and lower halves of the data, respectively as shown in this diagram:



Example 2.2.4

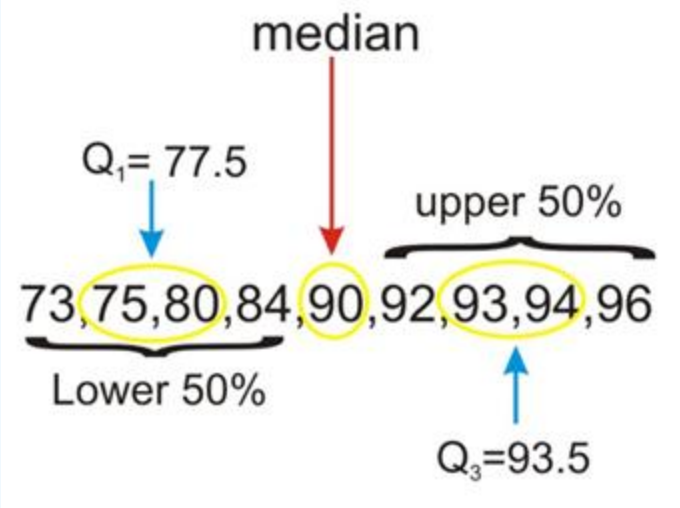
Find the quartiles of the data set:

73, 75, 80, 84, 90, 92, 93, 94, 96

Solution

There are $n = 9$ values in the data set, which is odd. Therefore, the value in the 5th position will be the median.

Here there are an even number of values remaining in the lower half and in the upper half. Follow the procedure for finding a median with an even number of data values and average the middle two values in each half as shown in the image below:



The quartiles are $Q_1 = 77.5$, $Q_2 = 90$, and $Q_3 = 93.5$.

You Try It $2.2.C$

A nutrition researcher records the daily grams of fiber consumed by 13 adults:

18 20 22 24 25 27 29 30 32 34 36 52 65

What are the quartiles?

Answer available at the end of this section.

Five-Number Summary

A useful way to summarize a set of quantitative data is to identify the values, once the data have been ordered, that divide the set into four equal sections. These values are the minimum, Q_1, the median, Q_3, and the maximum. Together, they form what is known as the five-number summary. This summary provides a concise description of the data set by highlighting its endpoints and key quartile positions.

The **five-number summary** of a dataset consists of its minimum, first quartile, median, third quartile, and maximum.

Example $2.2.5$

Find the five-number summary of the heights (in inches) of 9 females:

59 60 62 64 66 67 69 70 72

Solution

The data are already ordered. There is an odd number of data, $n = 9$. The median is located is at the 5th position, or 66.

59 60 62 64 **66** 67 69 70 72

To find Q_1 and Q_3, calculate the median of the lower half of the data and the median of the upper half of the data. Q_1 and Q_3 are each positioned between the 2nd and 3rd pieces of data in each half as shown:

59 **60 | 62** 64 **66** 67 **69 | 70** 72

$$Q_1 = \frac{60+62}{2} = 61 \text{ Median} = 66 \; Q_3 = \frac{69+70}{2} = 69.5$$

The minimum value is 67 and the maximum value is 72. We report the five-number summary as [59, 61, 66, 69.5, 72].

Example 2.2.6

A restaurant advertises that the typical number of French fries in a large order is 82. Roberta is a bit curious about this claim, so she bought a large order of fries each day for 18 days and counted the number of fries in the orders. Her data are below.

80 72 77 80 90 85 93 52 84 87 80 86 92 88 67 86 66 77

Find the five-number summary for the data in her sample.

Solution

First, order the data:

52 66 67 72 77 77 80 80 80

84 85 86 86 87 88 90 92 93

There $n = 18$ values, which is even. The median will be between the data values in positions #9 and #10:

52 66 67 72 77 77 80 80 **80**

84 85 86 86 87 88 90 92 93

$$\text{Median} = \frac{80+84}{2} = 82$$

There are nine values remaining in each half of the data set. Q_1 and Q_3 will each be at position #5 in each half:

52 66 67 72 **77** 77 80 80 **80**

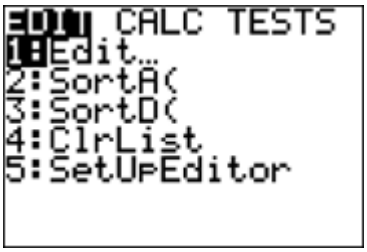
84 85 86 86 **87** 88 90 92 93

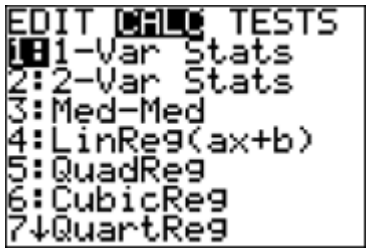
$$Q_1 = 77 \; Q_3 = 87$$

The minimum number of fries was 52 and the maximum number of fries was 93. The five-number summary is [52, 77, 82, 87, 93].

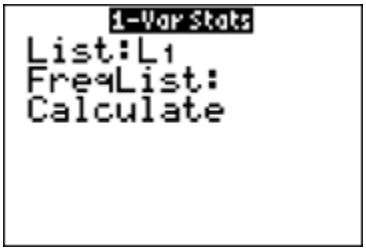
Five-Number Summary on the TI 83/84 Calculator

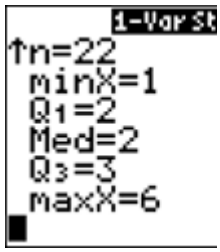
You may recall seeing the five-number summary when finding the mean and median on the calculator in the previous section. Follow that same procedure and make sure you "arrow d see the five-number summary. The sequence of screenshots is shown below for the data in 2.2.7.

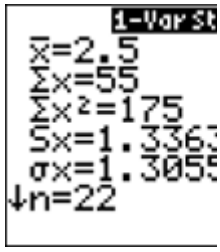












Test for Outliers using the Quartiles

Many data sets contain values that are either extremely large or extremely small compared to the rest of the observations. These unusual values are called **outliers**. If a data value lies far below or far above the general pattern of the distribution, it is considered an outlier. One common method for identifying outliers uses the quartiles from the **five-number summary**.

1.5IQR Outlier Test

1. Order the data and determine the Q_1 and Q_3.
2. Compute the **interquartile range (IQR)**, which is the difference $Q_3 - Q_1$.
3. Multiply the IQR by 1.5 to find the **step value**.
4. Determine the lower and upper fences by calculating $Q_1 - 1.5IQR$ and $Q_3 + 1.5IQR$.
5. Any data value that falls below the lower fence or above the upper fence is considered an outlier. Depending on the dataset, there may be none, one, or several outliers.

Example 2.2.7

Determine if any of the values in the French fries data from Example 2.2.7 were outliers.

Solution

Follow the steps above:

1. Recall from Example 2.2.7 that $Q_1 = 77$ and $Q_3 = 87$.
2. $\text{IQR} = Q_3 - Q_1 = 87 - 77 = 10$.
3. $\text{Step} = 1.5 \times \text{IQR} = 1.5 \times 10 = 15$.

4. Lower Frence $= Q_1 - \text{Step} = 77 - 15 = 62$ and Upper Frence $= Q_3 + \text{Step} = 87 + 15 = 102$.
5. Any values below 62 or above 102 will be considered outliers. Consulting the data, the day when Roberta got only 52 fries was an outlier. All other data lies in the range created by the fences [62,102].

Example 2.2.8

Based on past school spelling bee records, Ms. Lopez found the five-number summary of the number of words used before a winner was determined. Are there any outliers in the data?

Minimum = 48

Quartile 1 = 80

Median = 92

Quartile 3 = 99

Maximum = 135

Solution

1. $Q_1 = 80$ and $Q_3 = 99$.
2. $\text{IQR} = Q_3 - Q_1 = 99 - 80 = 19$.
3. $\text{Step} = 1.5 \times \text{IQR} = 1.5 \times 19 = 28.5$.
4. Lower Fence $= Q_1 - \text{Step} = 80 - 28.5 = 51.5$ and Upper Fence $= Q_3 + \text{Step} = 99 + 28.5 = 127.5$.
5. Any values outside the fenced interval will be considered outliers. Consulting the five-number summary, the minimum value was 48 so there are no outliers on the low side. The maximum value was 135 which is beyond the fence of 127.5. Therefore, the year when 135 words were used is an outlier. There *could* be other outliers on this higher end, but they cannot be identified without knowing the full set of raw data

You Try It 2.2.D

A fitness trainer tracks how many minutes clients spend on the treadmill during their weekly workout sessions. She summarizes the data with the following five-number summary: [12, 20, 28, 35, 55].

a. Find the lower fence and upper fence.
b. Is either the minimum or maximum value an outlier?

Answer available at the end of this section.

Section Summary

The pth percentile is the data value in an ordered set such that $p\%$ of the observations are less than or equal to that value and $(100 - p)\%$ are greater than it. Percentiles divide a dataset into meaningful positions that describe how individual values compare to the whole.

Percentiles are closely related to percentile rank, which describes the percentage of data values that fall at or below a particular observation. For example, if a value has a percentile rank of 68%, then 68% of the data are less than or equal to that value. Percentile rank focuses on the *relative standing* of a specific data point, while a percentile identifies the *cutoff value* itself.

Several important percentiles have special names:

- Quartile 1 Q_1 is the value at which 25% of the data lie at or below it.
- Median is the value at which 50% of the data lie at or below it.

- Quartile 3 Q_3 is the value at which 75% of the data lie at or below it.

These values form part of the five number summary, a concise description of a dataset consisting of the minimum, Q_1, median, Q_3, and maximum. Together, these five numbers divide the data into quarters and provide a clear picture of its spread and center.

The five-number summary is a set of five values that divides a data set into quarters: minimum, lower quartile, median, upper quartile, and maximum.

Quartiles are also used to check for outliers. The difference $Q_3 - Q_1$ is called the interquartile_range (IQR). Any data value that is less than $Q_1 - 1.5IQR$ or greater than $Q_3 + 1.5IQR$ is considered an outlier when compared to the rest of the dataset.

Section 2.2 Review Questions

1. Complete these statements about the meaning of a percentile.
   a. In general, _____% of the values in a data set fall at or below the 82nd percentile and _____% of the values in a data set fall above the 82nd percentile.
   b. If a sample contains 900 test scores, _____ of them would be equal to lower than the 22nd percentile.
   c. If a sample contains 900 test scores, _____ of them would be higher than the 68th percentile.

2. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is at the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

3. On a 20-question math test, the 70th percentile for number of correct answers was 16.
   a. What percentage of the test questions were answered correctly by a student whose score corresponds to the 70th percentile?
   b. What is the percentile rank of a student who answered 16 questions correctly?
   c. In this testing context, is a higher or lower percentile more desirable?

4. At the monthly 3rd grade teachers meeting, the principal tells you that the scores on the recent standardized test have been posted. Students can score anywhere from 100 to 400 on this test. The ordered raw scores for the $n = 30$ third graders in your class who took the test are shown below.

108	128	155	156	164
166	168	180	180	185
191	193	203	224	242
272	278	291	299	304
322	338	340	355	361
364	381	382	396	398

   a. Austin is one of your students and he scored "185" on the test. At what percentile rank did Austin score?
   b. What is the 60th percentile?

5. The numbers of coins that 12 randomly selected people had in their piggy banks are shown below. Find the five-number summary and test for outliers.

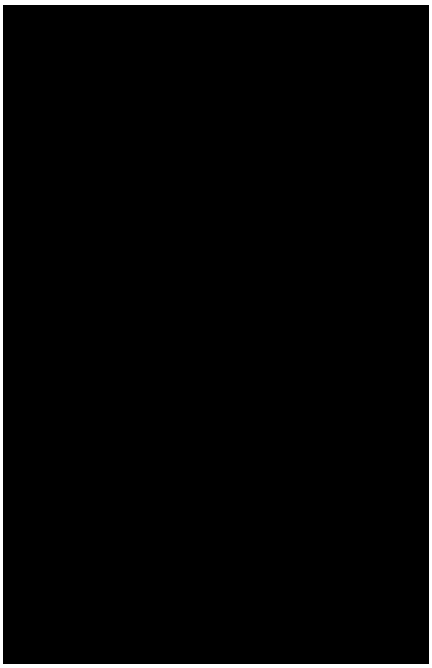
35	58	29	44	104	39
72	34	50	41	64	54

6. Use the five-number summary to determine whether the data set from which it was calculated contains any outliers.

Minimum = 5 Quartile 1 = 12 Median = 14 Quartile 3 = 16 Maximum = 24

7. The following data represent the average snowfall (in centimeters) for 18 Canadian cities for the month of January. Find the five-number summary and test for outliers.

Average Snow for Canadian Cities



8. Fitness Factory is a new gym that offers reasonably priced family packages. The following table represents the number of family packages sold during the opening month. Find the five-number summary and determine whether any values are outliers.

24	21	31	28	29
27	22	27	30	32
26	35	24	22	34
30	28	32	27	32
23	10	32	28	32
34	24	28	27	32

Answers

You Try It $2.2.A$

a. 85%
b. 15%

You Try It $2.2.B$

a. 23 cm
b. 30

You Try It $2.2.C$

|(Q_1 = 23\), $Q_2 \text{median} = 29$, and $Q_3 = 35$.

You Try It $2.2.D$

1. Lower Fence = -2.5 and Upper Fence = 57.5.
2. No. The minimum is greater than the lower fence. The maximum is less than the upper fence.

This page titled 2.2: Measures of Position is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

2.3: Measures of Spread

Learning Objectives

- Calculate measures of spread, including range, interquartile range, mean absolute deviation, and sample standard deviation.)
- Explain conceptual differences among measures of spread.
- Distinguish variance and standard deviation.
- Interpret differences in standard deviations.
- Compute a z-score.
- Interpret a z-score.

Introduction

In the previous two sections, we examined measures of center and position. Another important aspect of understanding a data set is the way its values are distributed or spread out. This characteristic is often described using the terms *variation* and *dispersion*. Measures of spread help us see how tightly the data cluster around the center or how widely they extend across the number line. In this lesson, we will investigate several commonly used measures of spread that provide deeper insight into the overall behavior of a data set.

Range

One basic measure of spread in a data set is the **range**. The range describes how widely the data values are dispersed by comparing the largest value to the smallest value. It is calculated by subtracting the minimum from the maximum:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Because it requires only two values, the range is quick to compute and provides an immediate sense of how far apart the data lie at the extremes. This makes it a useful first step in assessing variation. However, the range has limitations. Since it depends solely on the minimum and maximum, it does not reflect how the remaining data values are distributed. In addition, the range is not resistant to outliers. An unusually large or small value can greatly increase the range and give a misleading impression of the overall spread. For these reasons, the range is often used alongside other measures of dispersion.

Example 2.3.1

A teacher records the number of minutes eight students spent on homework last night. The data are listed in the order they were collected:

$$12 \;\; 5 \;\; 18 \;\; 9 \;\; 14 \;\; 7 \;\; 20 \;\; 11$$

Find the range.

Solution

To find the range, first identify the minimum and maximum values in the dataset.

- Minimum value: 4
- Maximum value: 18

The range is the difference between these two values:

$$\text{Range} = \text{Maximum} - \text{Minimum} = 18 - 4 = 14.$$

This tells us that the number of seeds eaten by the birds varies across a span of 14 seeds from the smallest to the largest observation. Because the range uses only the extreme values, it gives a quick snapshot of the overall spread but does not describe how the middle values behave. For example, most of the birds ate between 7 and 13 seeds, but the range is heavily influenced by the bird that ate 18 seeds.

Interquartile Range

Another important measure of spread is the **interquartile range**, commonly abbreviated **IQR**. The IQR describes how widely the middle half of the data is dispersed. It is calculated using the first quartile Q_1 and the third quartile Q_3:

$$\text{IQR} = Q_3 - Q_1.$$

Because it focuses on the central 50% of the distribution, the IQR provides a clearer picture of typical variation than the range which relies only on extreme values. No matter how the data are shaped, about half of all observations always fall between Q_1 and Q_3. A larger IQR indicates that the middle portion of the data is more spread out, while a smaller IQR suggests that the central values are more tightly clustered.

The IQR is also a resistant measure of spread, meaning it is not greatly affected by unusually large or small values. This makes it especially useful when a dataset contains outliers. For these reasons, the IQR is often used alongside the five-number summary to describe the overall behavior of a dataset.

Example 2.3.2

A researcher records the number of hours 10 students studied for an exam. The data, listed in the order collected, are

6 3 9 10 4 8 12 7 5 11

Find the IQR.

Solution

First, order the data and find the median:

3 4 5 6 **7 | 8** 9 10 11 12

$$\text{Median} = \frac{7+8}{2} = 7.5$$

Q_1 is the median of the lower half of the data:

3 4 **5** 6 7

$$Q_1 = 5$$

Q_3 is the median of the upper half of the data:

8 9 **10** 11 12

$$Q_3 = 10$$

$\text{IQR} = Q_3 - Q_1 = 10 - 5 = 5.$

The interquartile range is 5 hours, meaning the middle 50% of study times fall within a 5-hour span. Because half of all observations always lie between Q1 and Q3, this tells us how tightly the central portion of the data is clustered. In this example, the middle study times are moderately spread out.

You Try It 2.3.A

A meteorologist records the daytime high temperatures (in °F) for 9 cities on the same day. The temperatures, not yet ordered, are

68 75 71 80 77 69 83 72 78

Calculate the range and IQR.

Answer available at the end of this section.

Mean Absolute Deviation (MAD)

Another useful measure of spread is the **mean absolute deviation**, often abbreviated **MAD**. This statistic is commonly introduced in elementary and middle school because it provides an intuitive way to understand how far the data values vary from the mean.

Recall that the mean represents the numerical balancing point of a data set. To measure how the data are spread around this balancing point, we examine how far each individual value lies from the mean. The difference between a data value x_i and the mean $\bar{x}$ is called its **deviation**. Symbolically, the deviation is $x_i = \bar{x}$.

Let's take this simple data set of three randomly selected individuals' shoe sizes and explore how the MAD is calculated: 9.5 11.5 12

Step 1: Find the mean: $\bar{x} = \frac{9.5+11.5+12}{3} = 333 = 11$

Step 2: Find the deviation $x_i - \bar{x}$ for each data value.

For $x = 9.5$, $9.5 - 11 = -1.5$.

For $x = 11.5$, $11.5 - 11 = 0.5$.

For $x = 12$, $12 - 11 = 1$.

Notice that when a data value is less than the mean, its deviation from the mean is negative, and when a data value is greater than the mean, its deviation is positive. However, the very property that makes the mean the numerical balancing point of the data also makes it impossible to average these deviations directly. Because the mean balances the distribution, the sum of all deviations is always zero. Notice that $-1.5 + 0.5 + 1 = 0$.

To avoid this cancellation, we take the absolute value of each deviation before averaging. This produces the mean absolute deviation, a measure that reflects how far the data values typically lie from the mean without positive and negative deviations offsetting one another. Therefore, we need all the deviations to be positive before we add them up. One way to do this would be to make them positive by taking their absolute values.

Step 3: Take the absolute value of each deviation. MAD uses absolute deviations so negative values do not cancel positive ones.

$|-1.5| = 1.5$

$|0.5| = 0.5$

$|1| = 1$

Step 4: Find the mean of these absolute deviations.

$$\text{MAD} = \frac{1.5 + 0.5 + 1}{3} = 1$$

The mean absolute deviation of the shoe sizes is 1. This means that, on average, each shoe size differs from the mean size (11) by 1 unit.

Steps to Calculate the Mean Absolute Deviation, MAD

1. Find the mean, $\bar{x}$.
2. Find the deviation of each data value from the mean, $x_i - \bar{x}$.
3. Find the absolute value of each deviation, $|x_i - \bar{x}|$.
4. Find the mean of the absolute deviations.

As a formula, $\text{MAD} = \dfrac{\dsum|x_i - \bar{x}|}{n}$.

Example 2.3.3

Find the mean absolute deviation for these cell phone prices: $550, $320, $750, $500, $480.

Solution

Step 1: Find the mean: $\bar{x} = \frac{550+320+750+500+480}{5} = 26005 = 520$

Step 2: Find the deviation $x_i - \bar{x}$ for each data value.

For $x = 550$, $550 - 520 = 30$.

For $x = 320$, $320 - 520 = -200$.

For $x = 750$, $750 - 520 = 230$.

For $x = 500$, $500 - 520 = -20$.

For $x = 480$, $480 - 520 = -40$.

Note that $30 + (-200) + 230 + (-20) + (-40) = 0$.

Step 3: Take the absolute value of each deviation.

$|30| = 30$

$|-200| = 200$

$|230| = 230$

$|-20| = 20$

$|-40| = 40$

Step 4: Find the mean of these absolute deviations.

$$\text{MAD} = \frac{30 + 200 + 230 + 20 + 40}{5} = \frac{520}{5} = 104.$$

On average, the prices of smart phones in the data set are either $104 greater than or less than the mean price of a smart phone.

You Try It 2.3.B

Calculate the mean absolute deviation (MAD) of this data set: $\{4, 7, 10, 11, 18\}$.

Answer available at the end of this section.

Standard Deviation

Similarly, the **standard deviation** is a more advanced measure of spread used in higher-level statistics to describe the typical deviation of data values from their mean. Like the mean absolute deviation, it is based on how far each value lies from the mean. However, instead of taking absolute values of the deviations, the standard deviation is computed by **squaring** each deviation. Because the square of any real number is always positive or zero, squaring prevents positive and negative deviations from canceling each other out. This approach also gives greater weight to larger deviations, making the standard deviation especially useful for analyzing variability in more complex data sets.

Let's revisit the shoe size data we used before and calculate the standard deviation. Step 1 and Step 2 remain the same,

Step 3: Take the absolute value of each deviation. Standard deviation uses squared deviations so negative values do not cancel positive ones.

$$(-1.5)^2 = 2.25$$

$$(0.5)^2 = 0.25$$

$$(1)^2 = 1$$

Next, we want to find a special average of these squared deviations called the variance. Usually, to find an average, we divide by the number of terms in the sum. In the process of finding standard deviation, however, we divide by $n-1$. The variance of a sample is denoted by s^2.

Step 4: Find the variance:

$$s^2 = \frac{2.25 + 0.25 + 1}{3 - 1 = 2} = \frac{3.5}{2} = 1.75.$$

Step 5: The square root of the variance is the standard deviation, s.

$$s = \sqrt{1.75} \approx 1.32$$

The standard deviation indicates how far, on average, the data values lie from the mean. In this sample, a standard deviation of approximately 1.32 means that a typical data value is about 1.32 units away from the mean. It is a measure of how tightly or loosely the data are clustered around the mean.

The standard deviation is always positive or zero. It is small when the data values are concentrated close to the mean, showing little variation. It becomes larger when the data values are more widely spread out from the mean, reflecting greater variability in the dataset.

Steps to Calculate the Sample Standard Deviation, s

1. Find the mean, $\bar{x}$.
2. Find the deviation of each data value from the mean, $x_i - \bar{x}$.
3. Square each deviation, $(x_i - \bar{x})^2$.
4. Add the squared deviations and divide by $n-1$ to find the variance, s^2.
5. Find the standard deviation, s by taking the square root of the variance.

The formulas for the variance and standard deviation are

$$s^2 = \frac{\dsum(x_i - \bar{x})^2}{n-1} \text{ and } s = \sqrt{\frac{\dsum(x_i - \bar{x})^2}{n-1}}.$$

Example 2.3.4

The junior high school baseball team has played 6 games. Find the standard deviation for the number of runs scored: 7, 0, 6, 8, 5, 4.

Solution

Step 1: Find the mean: $\bar{x} = \frac{7+0+6+8+5+4}{6} = 306 = 5$

Step 2: Find the deviation $x_i - \bar{x}$ for each data value.

For $x = 7$, $7 - 5 = 2$.

For $x = 0$, $0 - 5 = -5$.

For $x = 6$, $6 - 5 = 1$.

For $x = 8$, $8 - 5 = 3$.

For $x = 5$, $5 - 5 = 0$.

For $x = 4$, $4 - 5 = -1$.

Note that $2 + (-5) + 1 + 3 + 0 + (-1) = 0$.

Step 3: Square each deviation.

$(2)^2 = 4$

$(-5)^2 = 25$

$(1)^2 = 1$

$(3)^2 = 9$

$(0)^2| = 0$

$(-1)^2 = 1$

Step 4: Find the variance:

$$s^2 = \frac{4 + 25 + 1 + 9 + 0 + 1}{6 - 1} = \frac{40}{5} = 8.$$

Step 5: Find the square root of the variance.

$s = \sqrt{8} \approx 2.83$

The number of runs the team scores in a typical game tends to differ from the mean of 5 runs by about 2.83 runs. This indicates moderate variability. The team's scoring is not extremely consistent, but not wildly unpredictable either.

You Try It 2.3.C

Calculate the standard deviation of this sample data set: $\{4, 7, 10, 11, 18\}$.

Answer available at the end of this section.

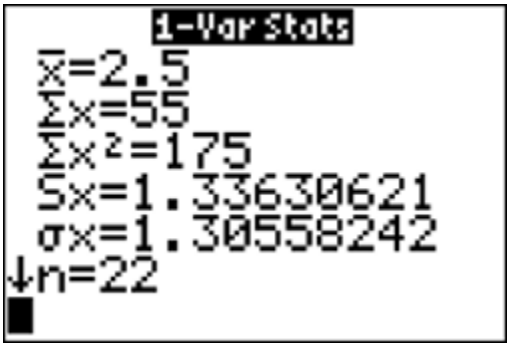
When we calculate the sample standard deviation, we divide by $n - 1$ rather than n. This adjustment is called *Bessel's correction,*

and it exists because a sample does not perfectly represent the entire population. Using $n-1$ corrects for a built-in bias that occurs when we estimate variability from a sample. A sample mean $barx$ is calculated from the sample itself, so it tends to sit closer to the sample data than the true population mean mu would. Because the deviations $x_i - \bar{x}$ are measured from the sample mean, they are slightly smaller than the deviations from the population mean. This makes the raw average of squared deviations too small—it underestimates the true spread of the population. Dividing by $n-1$ instead of n increases the value just enough to correct this underestimation.

To find standard deviation of population data rather than sample data, Bessel's correction is not needed and so the formula for standard deviation would be $\sigma = \sqrt{\frac{\dsum(x_i-\mu)^2}{N}}$.

Finding Standard Deviation on the TI 83/84 Calculator

The sample standard deviation of data input into L1 is denoted on the 1-VarStats screen as s_x. To find the variance, you can square this value. If the data in L1 is population data instead of sample data, then the standard deviation of the population is denoted by σ_x.



Using Standard Deviation to Measure Unusualness

Standard deviation is not only a measure of spread—it is also a powerful tool for identifying unusual or unexpected data values. Because the standard deviation describes how far typical values fall from the mean, we can use it as a kind of “yardstick” for measuring how far any individual value lies from the center of the distribution.

When a data value is within one standard deviation of the mean, it is generally considered pretty typical. Values that fall two or more standard deviations away from the mean are often considered unusual, because they lie far outside the range where most data tend to cluster. This idea becomes especially useful when comparing values from different datasets or when determining whether a particular observation stands out.

z-scores

To make this idea more precise, statisticians use a standardized measure called a **z-score**. A z-score tells how many standard deviations a data value is above or below the mean. It converts raw data into a common scale, making it easier to compare values across different contexts.

To compute a z-score, we compare a data value x to the mean $\bar{x}$ and then scale that difference by the standard deviation s.

> A **z-score** is a standardized measure that tells how mana data may standard deviations a data value lies above or below the mean. Symbolically, the z-score for a data value in sample x is
>
> $$z = \frac{x - \bar{x}}{s}.$$

A positive z-score indicates that the value lies above the mean, while a negative z-score indicates that it lies below the mean. The magnitude of the z-score shows how far the value is from the mean relative to the overall spread of the data.

A z-score allows us to classify values in a data set as typical or unusual:

- A z-score near 0 represents a value close to the mean.
- z-scores between –1 and 1 usually indicate typical values within one standard deviation of the mean.
- z-scores beyond ±2 often signal unusual or extreme observations.
- Very large positive or negative z-scores identify values that lie far from the center of the distribution.

Because z-scores place all data on the same standardized scale, they make it possible to compare values from different datasets, different units, or different contexts. This standardization becomes especially important in later topics such as the normal distribution.

Example 2.3.5

A grocery store tracks the weights of cartons of strawberries. In a recent sample, the mean carton weight was 14 ounces, with a standard deviation of 2 ounces.

One carton weighed 15 ounces. Find the z-score for this carton.

Solution

$$z = \frac{x - \bar{x}}{s} = \frac{15 - 14}{2} = \frac{1}{2} = 0.5$$

The z-score is 0.5 which means that this 15-ounce carton is only $\frac{1}{2}$ of a standard deviation above the mean.

Based on the rule of thumb for z-scores, this carton is pretty typical since its z-score falls between -1 and 1.

Example 2.3.6

A sample of books in the library has mean length of 350 pages with a standard deviation of 100 pages. Is a book with 125 pages unusually short?

Solution

Find the *z*-score of the 125-page book to see how many standard deviations this book lies from the mean:

$$z = \frac{x - \bar{x}}{s} = \frac{125 - 350}{100} = \frac{-225}{100} = -2.25$$

The 125-page book is more than two standard deviations below the mean so based on the rule of thumb for z-scores, this book is considered far below average and unusual.

You Try It 2.3.D

A sample of commuter train trips has a mean travel time of 50 minutes with a standard deviation of 12 minutes.

Find the z-score for a trip lasting 68 minutes. Is this trip unusually long?

Answer available at the end of this section.

Using z-scores to Compare Data Values

When two data sets have different means and standard_deviations, comparing raw scores directly can be misleading. A score that looks high in one class may not be as impressive if the class tends to score well overall, while a lower raw score in another class may actually represent stronger performance if the class average is lower. To make fair comparisons, we convert raw scores into z-scores, which measure how many standard deviations a value lies above or below its mean.

The z-scores place different data sets on the same standardized scale, allowing meaningful comparisons across different subjects, tests, or contexts.

Example 2.3.7

John is taking biology and math. He took an exam in both classes. The class means and standard deviations were

- In biology, the class mean was 82 and the standard deviation was 6.
- In math, the class mean was 78 and the standard deviation was 3.

John earned 85 in biology and 80 in math. Which score is more impressive relative to the rest of the class?

Solution

Although John's raw score of 85 in biology is higher than his 80 in math, raw scores alone do not tell us how he performed compared with his classmates. To compare his performance fairly, we compute z-scores for both classes.

$$z = \frac{x - \bar{x}}{s}$$

$$z_{\text{biology}} = \frac{85 - 82}{6} = \frac{3}{6} = 0.5$$

$$z_{\text{math}} = \frac{80 - 78}{3} = \frac{2}{3} \approx 0.67$$

John's biology score is 0.5 standard deviations above the mean, while his math score is about 0.67 standard deviations above the mean. Since higher z-scores represent more impressive performance relative to the group, John's 80 in math is more impressive than his 85 in biology when considered in the context of each class's overall performance.

Example 2.3.8

A wildlife biologist collects sample data on two kinds of animals.

- From a sample of adult rabbits, the biologist finds a mean weight of $\bar{x} = 2.4$ kg and a standard deviation of $s = 0.3$ kg. A particular rabbit in the study weighs 3.0 kg.
- From a sample of adult foxes, the biologist finds a mean weight of $\bar{x} = 6.8$ kg and a standard deviation of $s = 0.5$ kg. A particular fox in the study weighs 5.6 kg.

Determine which animal's weight is **more unusual** relative to its sample.

Solution

$$z = \frac{x - \bar{x}}{s}$$

$$z_{\text{Rabbit}} = \frac{3.0 - 2.4}{0.3} = \frac{0.6}{0.3} = 2$$

$$z_{\text{Fox}} = \frac{5.6 - 6.8}{0.5} = \frac{-1.2}{0.5} = -2.4$$

The rabbit's weight is 2 standard deviations *above the mean*.The fox's weight is 2.4 sample standard deviations *below the mean*. To compare unusualness, compare their magnitudes, or absolute values.

$$|z_{\text{Rabbit}}| = |2| = 2$$

$$|z_{\text{Fox}}| = |-2.4| = 2.4$$

Since $2.4 > 2$, the fox's weight is more unusual relative to its species.

You Try It $2.3.E$

A nutrition researcher analyzes the sodium content of several packaged foods and computes their z-scores:

- Food A: $z = -2.1$
- Food B: $z = 1.2$
- Food C: $z = -0.4$

Place the foods in order from least unusual to most unusual based on their z-scores.

Answer available at the end of this section.

Using a z-score to Find a Value

A z-score tells us how many standard deviations a data value lies above or below the sample mean. Sometimes we know the z-score, the sample mean, and the sample standard deviation, and we want to determine the original data value. To do this, we can rearrange the z-score formula to solve for x:

$$z = \frac{x - \bar{x}}{s}$$

Multiply both sides of the formula by s and add $\bar{x}$.

$$x = \bar{x} + zs$$

Example $2.3.9$

A meteorologist analyzes a sample of daily wind speeds (in miles per hour). In his sample, the mean wind speed was 12.5 mph and the standard deviation was 3.2 mph.

On one particular day, the wind speed has a z-score of $z = .75$. What was the wind speed on that day?

Solution

A z-score of 0.75 means that on this day, the wind speed was 0.75 standard deviations above the mean.

Use the rearranged formula for a z-score and substitute $\bar{x} = 12.5$, $s = 3.2$ and $z = 0.75$.

$$x = \bar{x} + zs$$

$$x = 12.5 + (0.75)(3.2)$$

$$x = 12.5 + 2.4$$

$$x = 14.9$$

The wind speed on this day was 14.9 mph.

You Try It $2.3.F$

A grocery analyst studies a sample of weekly spending totals for customers at a local supermarket. In the sample, the mean weekly spending was \$82, and the standard deviation was \$15.

One customer's weekly spending had a z-score of $z = -0.8$. What was this customer's spending for the week?

Answer available at the end of this section.

Section Summary

Measures of spread describe how much variability appears in a set of data. They indicate how consistent or inconsistent the data values are.

The range, computed as "maximum – minimum," is a simple measure of spread that shows how far apart the smallest and largest values are. However, it ignores all data values between these two extremes.

The interquartile range (IQR) is used when the median is the chosen measure of center. The IQR is the difference between Quartile 3 and Quartile 1: $\text{IQR} = Q_3 - Q_1$. It measures the spread of the middle 50% of the data.

The mean absolute deviation (MAD) measures spread from the mean by averaging the absolute deviations of all data values from the mean. Its formula is $\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$.

The standard deviation describes the typical distance of data values from the mean. It is found by averaging the squared deviations to obtain the variance, then taking the square root. In more advanced statistics, this is the preferred measure of spread from the mean. The formula for the sample standard deviation is $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$.

A z-score measures how many standard deviations a data value lies above or below the mean. Its formula is $z = \frac{x - \bar{x}}{s}$. Z-scores help determine whether a data value is far above average or far below average. Values with $z > 2$ or $z < -2$ are typically considered unusual. Z-scores are also useful for comparing data values from different data sets because they place all values on a common standardized scale.

Section 2.3 Review Questions

1. The following data represent the height, in feet, of various buildings in the city: 54, 56, 63, 65, 62, 59, 56, 58
   a. Find the range.
   b. Find the interquartile range (IQR).
   c. Find the mean absolute deviation (MAD.)

2. Use the rainfall data in the table below to answer this question.
   a. Calculate the sample mean.

b. Complete the chart to calculate the variance and the standard deviation.
c. Confirm your answer to part b using technology.

Problem 2

Year	Rainfall (inches) x	Deviation $x - \bar{x}$	Squared Deviations $(x - \bar{x})^2$
1998	90		
1999	56		
2000	60		
2001	59		
2002	74		
2003	76		
2004	81		
2005	91		
2006	47		
2007	59		

3. According to the U.S. Department of Agriculture, ten to twenty earthworms per cubic foot is a sign of healthy soil. Mr. Green checked the soil in his garden by digging 7 one-cubic-foot holes and counting the earthworms. The table shows the number of earthworms he found in each hole.

[redacted]

a. Find the range.
b. Find the interquartile range (IQR.)
c. Find the mean absolute deviation (MAD) by hand.
d. Find the sample variance and sample standard deviation by hand.
e. Confirm your answers to part (d) using technology.
f. In how many holes was the number of earthworms more than 1 standard deviation away from the mean?
g. In how many holes was the number of earthworms more than 2 standard deviations away from the mean?
h. Find the *z*-score of Hole G.

4. The mean speed of a sample of vehicles along a stretch of highway is 56 mph with a standard deviation of 4 mph. Find the *z*-scores of each of these speeds.
   a. 62 mph
   b. 47 mph
   c. 56 mph

5. A leg bone in dinosaur fossils has a mean length of 5 feet with a standard deviation of 3 inches. What is the *z*-score that corresponds to a length of 62 inches? (*Caution: Make sure you use consistent units.*)

6. Sampson recently took a quiz in biology and a midterm exam in marketing.
   a. In biology, the class mean on the quiz was 15 points with a standard deviation of 3.6 points. Calculate Sampson's z-score (to 2 decimal places) if he earned a score of 24.
   b. In marketing, the class mean on the midterm was 75% with a standard deviation of 4.3%. Calculate Sampson's z-score (to 2 decimal places) if he scored 84%.
   c. In which class did Sampson have a more impressive performance relative to the rest of the class? Justify your answer.

7. A town's January high temperatures averages 36°F with a standard deviation of 10°, while in July the mean high temperature is 74° and standard deviation is 8°. In which month is it more unusual to have a day with a high temperature of 55°?

8. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100.
   a. Which cost is the lowest, when compared to other instruments of the same type? Justify your answer.
   b. Which cost is the highest when compared to other instruments of the same type? Justify your answer.

9. On the first statistics test in Ms. Upshaw's class, the mean was 73.5 and the standard deviation was 12.
   a. If Ben had a z-score of 1.56 on this test, what was his original raw test score?
   b. If Susan had a z-score of -2.08 on this test, what was her original raw test score?

Answers

You Try It $2.3.A$

range = 15°F and IQR = 9°F

You Try It $2.3.B$

3.6

You Try It $2.3.C$

$s \approx 5.244$

You Try It $2.3.D$

The z-score is 1.5. While it might not be typical, this trip is not considered unusually long either since z is not beyond ±2.

You Try It $2.3.E$

Food C, Food B, Food A

You Try It $2.3.F$

$70

This page titled 2.3: Measures of Spread is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

2.4: Changes to a Data Set

Learning Objectives

- Describe effects of adding or subtracting a constant to all values in a data set.
- Describe effects of multiplying or dividing all values in a data set by a constant.
- Describe the effect of changing an extreme value within a data set.
- Tell whether measures are resistant or non-resistant.

Introduction

In this section, we explore how measures of center, position, and spread respond when changes are made to a data set. We will examine two types of changes: altering every value in the data set through addition and subtraction or multiplication and division, and altering only a single value within the set. Understanding how these transformations affect the mean, median, quartiles, range, and standard deviation helps us predict how a data set behaves when it is shifted or rescaled, and how sensitive each measure is to individual data values.

Shifting the Data: Adding or Subtracting a Constant

In this section, we examine what happens to measures of center, position, and spread when the same constant is added to (or subtracted from) every value in a data set. This type of transformation **shifts** the entire data set left or right on a number line without changing the spacing between the values. To see how this affects different statistical measures, we explore a simple scenario:

Ms. Maple gives a science quiz to 6 students. Their scores (out of 40 points) are

$$20\ 28\ 32\ 33\ 35\ 38.$$

This table gives some measures of center, position, and spread for this data set:

Summary Statistics for Original Quiz Scores

Center	$\bar{x} = \frac{20+28+32+33+35+38}{6} = \frac{186}{6} = 31$ $\text{Median} = \frac{32+33}{2} = 32.5$
Position	$\text{Five-number Summary: } [20, 28, 32.5, 35, 38]$
Spread	$\text{Range: } = 38 - 20 = 18$ $\text{IQR: } = 35 - 28 = 7$ $\text{MAD: } = \frac{28}{6} \approx 4.67$ $s \approx 6.45$

Example 2.4.1

Ms. Maples thinks that the quiz should be worth 50 points instead of 40 points. She decides to add 10 points to every score. How does this change affect the summary statistics for the group of scores?

Solution

Here's a table of the transformed data. Notice that each score has been increased by 10 points.

Original Quiz Scores Increased by 10

x	20	28	32	33	35	38
$x + 10$	30	38	42	43	45	48

Summary Statistics for Original Quiz Scores Increased by 10

Center	$\bar{x} = \frac{30 + 38 + 42 + 43 + 45 + 48}{6} = \frac{246}{6} = 41$ $\text{Median} = \frac{42 + 43}{2} = 42.5$
Position	$\text{Five-number Summary: } [30, 38, 42.5, 45, 48]$
Spread	$\text{Range: } = 48 - 30 = 18$ $\text{IQR: } = 45 - 38 = 7$ $\text{MAD: } = \frac{28}{6} \approx 4.67$ $s \approx 6.4$

Comparing the summary statistics for the adjusted scores with those from the original data set, we see that adding 10 points to every score increases each measure of center and position by exactly 10 points. Both the mean and median increase by 10, and every value in the five-number summary shifts higher by 10 as well. This behavior is expected: when every data value is shifted to the right on a number line by the same amount, all measures that depend on center or position of the data shift by that same amount.

In contrast, the measures of spread -- the range, IQR, MAD, and sample standard deviation -- remain unchanged. Shifting every data value by the same constant does not make the data more or less spread out. Because all values move together, the distances between the data values stay exactly the same. Spread measures, which depend on those distances, do not change.

These patterns always hold: adding or subtracting a constant affects measures of center and position but does not affect measures of spread.

Rescaling the Data: Multiplying or Dividing by a Constant

What happens to the measures of center, position, and spread when we *multiply* the same value by every value in the data set? This type of transformation **scales** the entire data set. Let's explore this idea by adjusting the original quiz scores again.

Example 2.4.2

Ms. Maple thinks that the quiz should be worth 80 points instead of 40 points and decides to double every score. How does multiplying by 2 affect the summary statistics for the group of scores?

Solution

Original Quiz Scores Multiplied by 2

x	20	28	32	33	35	38
$2 \cdot x$	40	56	64	66	70	76

Summary Statistics for Original Quiz Scores Multiplied by 2

Center	$\bar{x} = \frac{40 + 56 + 64 + 66 + 70 + 76}{6} = \frac{372}{6} = 62$ $\text{Median} = \frac{64 + 66}{2} = 65$
Position	$\text{Five-number Summary: } [40, 56, 65, 70, 76]$
Spread	$\text{Range: } = 76 - 40 = 36$ $\text{IQR: } = 70 - 56 = 14$ $\text{MAD: } = \frac{56}{6} \approx 9.33$ $s \approx 12.9$

Comparing the two tables, we see that multiplying each individual score by 2 doubles the measures of center -- the mean and median -- as well as the measures of position, such as the five-number summary. This mirrors what happened when a constant was added: any measure that depends on the location of the data is scaled by the same factor applied to every data value.

However, in contrast to addition, the measures of spread -- the range, IQR, MAD, and sample standard deviation -- also double. This occurs because multiplying the data values stretches the distances between them. For example, doubling 20 increases it by 20 points, while doubling 38 increases it by 38 points. Larger values move farther than smaller values, so the overall spread widens. The spread measures reflect this increased distance among the data points.

This pattern will always hold. Multiplying every value in a data set by a constant multiplies all measures of center, position, and spread by that same constant. The same is true when dividing every data value by a constant: all measures of center, position, and spread are divided by that constant as well.

Example 2.4.3

A sample of fitness trackers sold at TrackTime Electronics was collected. Summary statistics for the prices (in dollars) are shown below.

Prices for a Sample of Fitness Trackers

Statistic	Value
Mean	\$129.50
Standard deviation	\$42.30
Median	\$118.00
Q_3	\$155.00
Range	\$160.00

TrackTime Electronics decides to offer a promotion by reducing the price of every fitness tracker by \$25. Find the new value of each statistic after the discount is applied.

Solution

Subtracting a constant from every data value shifts the entire set of data lower by that constant. Measures of center and location (mean, median, quartiles) decrease by the same amount, while measures of spread (range, IQR, MAD, standard deviation) remain unchanged.

Prices for a Sample of Fitness Trackers

Statistic	Value	Value After Promotion Is Applied
Mean	\$129.50	Mean changes by the same amount. The new mean is $\$129.50 - \$25 = \$104.50$.
Standard deviation	\$42.30	Standard deviation does not change when every value is shifted. The new standard deviation is \$42.30.
Median	\$118.00	Median changes by the same amount. The new median is $\$118.00 - \$25 = \$93.00$.
Q_3	\$155.00	Q_3 changes by the same amount. The new Q_3 is $\$155.00 - \$25 = \$130.00$
Range	\$160.00	Range does not change when every value is shifted. The new range is \$160.00.

You Try It 2.4.A

A greenhouse records the heights of a sample of houseplants. The mean is 42.0 cm with standard deviation 6.5 cm. The median is 40.0 cm with IQR 15 cm.

After a month of consistent care, every plant grows exactly 3.2 cm. Find the new value of each statistic after the growth occurred.

Answer available at the end of this section.

You Try It $2.4.B$

The mean and standard deviation for a data set are 65 and 12.5, respectively.

If all values in the data set are doubled, what is the mean and standard deviation?

Answer available at the end of this section.

Example $2.4.4$

On a 40-question multiple choice exam, a teacher records the number of problems each student answered correctly. From the class results, the teacher calculates the

- Mean: 28.7 problems
- Median: 29 problems
- Standard deviation: 4.3 problems

To determine the exam grade, the teacher uses the following rule:

- Each correctly answered problem is worth 3 points.
- Every student receives an additional 5 points for writing their name on the exam.

That is, if a student answered x problems correctly, their grade G is

$$G = 3x + 5.$$

Find the mean, median, and standard deviation of the exam grades.

Solution

Two separate transformations are applied to the original data set of problems correct. First, each number of correct problems is multiplied by 3 (a scale). Then, 5 points are added (a shift).

Remember:

- Measures of center and location (mean, median) are affected by both a scale and a shift.
- Measures of spread (standard deviation) are affected only by a scale.

The mean changes in the same way when the data is scaled by 3 and then shifted by 5. The mean of the grades is $3(28.7) + 5 = 86.1 + 5 = 91.1$.

The median changes in the same way both when the data is scaled by 3 and shifted by 5. The median of the grades is $3(29) + 5 = 87 + 5 = 92$.

The standard deviation changes in the same way when the data is scaled by 3, but it does not change when the data is shifted by 5. The standard deviation of the grades is $3(4.3) = 12.9$.

Effect of Adding/Removing a Data Value on the Measures of Center

Recall that one way to understand the mean is to think of it as a balancing point of the data set. Each data value contributes weight to one side or the other, and the mean is the point at which the data would balance if placed on a number line. Because of this, adding a new data value—or removing an existing one—will generally shift the balancing point unless the value happens to lie exactly at the mean.

When a new data value is added above the mean, the balance shifts to the right, causing the mean to increase. Likewise, removing

a data value that is below the mean also causes the mean to increase, because the remaining values are, on average, larger. Conversely, adding a data value below the mean or removing a data value above the mean will cause the mean to decrease. In short, the mean moves in the direction of the new imbalance created by adding or removing a value.

The median *can* also be affected when a data value is added or removed, but not always. Because the median depends only on the position of values rather than their magnitudes, adding or removing a value may shift the middle position—or it may leave the middle position unchanged. For example, adding a value far above or far below the existing data might not change which value falls in the middle, while adding a value near the center could shift the median.

In summary, adding or removing a single data value can change both the mean and the median, but the mean is always affected unless the new value equals the mean, while the median is affected only when the middle position of the ordered data changes.

Example 2.4.5

In the data set {1, 2, 3, 4, 4, 6, 6}, the median is 4.

What happens to the median if the value 3 is removed? What if one of the 6's is removed?

Solution

If the value 3 is removed, the median of {1, 2, 4, 4, 6, 6} is still 4. It's unchanged.

If the value 6 is removed, the median of {1, 2, 3, 4, 4, 6} is now 3.5. It changes.

In either case, the median does not change dramatically.

It is also important to recognize that adding or removing an extreme value—a value far from the rest of the data—will affect the mean much more than the median. The mean is sensitive to the magnitude of every data value because it incorporates all values directly in its calculation. In contrast, the median depends only on the position of values in the ordered list, not on how large or small those values are. As a result, extreme values can pull the mean noticeably upward or downward, while the median often remains unchanged or changes only slightly.

Let's explore an easy example to see this difference in action.

Example 2.4.6

Consider the very simple data set {1, 2, 3}. The mean is 2 and the median is 2. Suppose we include a huge value in the data set, like 100, so that the new data set is {1, 2, 3, 100}.

What happens to the median and mean?

Solution

The median increases only slightly—from 2 to 2.5—because the median depends on position, not magnitude. The extreme value 100 is placed at the end of the ordered list and does not strongly influence the middle position.

The mean becomes

$$\bar{x} = \frac{1+2+3+100}{4} = \frac{106}{4} = 26.5.$$

The mean increases dramatically -- from 2 to 26.5 -- because the mean incorporates all data values directly. The extreme value 100 heavily pulls the mean upward.

Example 2.4.7

Four friends play a round of golf and their scores are {70, 71, 71, 103}. What effect does removing the 103 have on the mean and median of the set?

Solution

Removing the score 103 has no effect on the median -- it stays at 71. This makes sense because the median depends only on the center position of the ordered data, and the extreme value does not influence which value lies in the middle.

However, removing 103 has a large effect on the mean. The mean drops from 78.75 to 70.67, a change of more than eight points. This happens because the mean incorporates all data values, and the extreme score of 103 pulls the mean upward. Once that value is removed, the mean falls back to a level that more accurately reflects the tightly clustered scores.

This example illustrates that the mean is highly sensitive to extreme values, while the median is much more resistant to them.

Example 2.4.8

Chuck selected 12 grocery items, each with a different price. The mean cost of the items was $7.41, and the median cost was $5.35. One of the items was a prime rib priced at $39.99. Chuck then decided to put the prime rib back and purchase only the remaining 11 items.

Describe how removing the prime rib affects the mean and the median of the prices.

Solution

Removing an item that is both extremely large and greater than the mean will cause the mean to decrease. The total cost decreases by $39.99 while the number of items decreases by only one, so the mean of the remaining 11 items must be lower.

The median also decreases, but only slightly. With 12 items, the median is the average of the 6th and 7th prices when ordered. After removing the largest item, the data set has 11 items, so the median becomes the 6th price alone. Because the removed value was the maximum, the middle of the data set shifts downward.

Thus, both the mean and the median decrease, but the mean decreases more noticeably.

Resistant vs. Non-resistant Measures

Statistical measures that change noticeably in the presence of extreme values or outliers are called **non-resistant measures.**

The mean is a non-resistant measure of center because a single unusually large or small value can pull it upward or downward. In contrast, the median is largely unaffected by outliers, making it a **resistant measure** of center.

Measures of spread behave similarly. The interquartile range (IQR) is resistant because it is computed using only the first and third quartiles; extreme values do not influence its calculation. However, the range, mean absolute deviation (MAD), and standard deviation all incorporate the most extreme observations and are therefore non-resistant measures of spread.

When selecting appropriate measures of center and spread, it is essential to consider whether outliers are present. If the data contain extreme values, the median and IQR provide more reliable descriptions of the distribution. In contrast, the mean, range, MAD, and standard deviation may be overly influenced by outliers and fail to represent the overall pattern of the data.

Example 2.4.9

A group of 9 friends likes to bowl together, and each friend keeps track of their all-time highest score in a single game. Eight of the friends have high scores between 140 and 180. Angie's original high score is 195, which is higher than everyone else's. Later, Angie bowls an even better game and raises her all-time high score to 240 (replacing the 195).

Describe how Angie's new high score will affect the median, mean, range, Q_3, IQR, and standard deviation of the 9 high scores.

Solution

- Median: With 9 scores, the median is the 5th value in the ordered list. Changing only the largest value does not change the 5th value, so the median stays the same.
- Mean: The mean is the sum of all 9 scores divided by 9. Angie's high score increases from 195 to 240, raising the total by $240 - 195 = 45$ points. Since the total increases while the number of scores stays 9, the mean increases.
- Range: The range is the difference between the maximum and minimum. The minimum score does not change, but the maximum increases from 195 to 240. Therefore, the range increases by 45.
- Q_3: For 9 ordered values, the upper half consists of the 4 largest values (positions #6–9). Angie's score is the 9th (largest) value. Q_3 is the median of the upper half, which depends on the middle of positions #6–9, not on the largest value alone. Since only the largest value changes, Q_3 stays the same.
- IQR: The interquartile range is IQR=Q_3–Q_1. We have already seen that Q_3 does not change, and none of the lower scores (which determine Q_1) change either. Thus, the IQR stays the same.
- Standard deviation: Standard deviation measures how far the scores typically are from the mean. Making the largest score even larger increases the spread of the data and pulls the mean upward. Angie's new score is farther from the mean than her old score, so the standard deviation increases.

You Try It 2.4.C

A teacher records the heights of 9 children in her class. Eight of the children have heights between 118 cm and 135 cm. Leo's height is originally 108 cm, which is lower than everyone else's.

Later in the year, Leo grows to 112 cm, replacing the old minimum value of 108 cm. All other heights remain the same

Tell whether each of these statistics for the heights of the 9 children increases, decreases, or remains the same.

a. mean
b. median
c. range
d. Q_1
e. IQR
f. standard deviation

Answer available at the end of this section.

Section Summary

Adding or subtracting a constant from every value in a data set shifts the entire distribution without changing its spread. The mean, median, and the five-number summary all increase or decrease by exactly that constant. Measures of spread -- such as the range, IQR, MAD, and standard deviation -- remain unchanged because the distances between data points stay the same.

Multiplying or dividing every data value by a constant scales the entire distribution. In this case, all numerical measures -- including the mean, median, five-number summary, range, IQR, MAD, and standard deviation -- are multiplied or divided by that

constant.

Some measures respond strongly to extreme values, while others do not. Resistant measures, such as the median and IQR, are largely unaffected by outliers and provide reliable descriptions of center and spread when the data contain extreme values. Non-resistant measures, including the mean, range, MAD, and standard deviation, can be distorted by unusually large or small observations. When a data set includes outliers, choosing resistant measures—specifically the median for center and the IQR for spread—gives a more accurate and meaningful summary of the distribution.

Section 2.4 Review Questions

1. Susan owns a small business that employs 5 other people. Susan makes $80,000 and pays the other 5 employees between $40,000 and $50,000 a year.
   a. If Susan gives everyone a $2000 raise (including herself), describe how this will affect the mean, median, IQR, and standard deviation of the salaries.
   b. If Susan gives everyone a 10% raise (including herself), describe how this will affect the mean, median, IQR, and standard deviation of the salaries. (*Hint: Each new salary will be 110% of the old salary.*)
   c. If Susan gives only herself a $10,000 raise, describe how this will affect the mean, median, IQR, and standard deviation of the deviation of the salaries.

2. A café tracks the weights of individual bags of roasted coffee beans. From a recent sample, the mean bag weight is 6.75 pounds, and the standard deviation is 1.25 pounds.The café decides to include an additional 2.3-pound promotional sampler inside every bag. What are the new mean and new standard deviation of the bag weights?

3. A bakery measures the weights of individual bags of flour and finds that the mean weight is 6.75 pounds with a standard deviation of 1.25 pounds. Since 1 pound = 16 ounces, what are the mean and standard deviation of the flour bag weights in ounces?

4. Sam played 5 rounds of golf and his lowest score on any round was 78. Later it was discovered that Sam broke a rule while playing so that score was removed from the data set. Describe the effect of removing the score of 78 on the mean and median.

5. In Lois' 2nd grade class, all the students were between 45 and 52 inches in height. One day Lucas, who is 62 inches tall, joined the class. Describe the effect of Lucas joining the class on the mean height and median height.

6. Mr. Franklin was interested his average monthly electric bill. He gathered his bills for the past 12 months and calculated some descriptive statistics.

Descriptive Statistics for Mr. Franklin's Monthly Electric Bills

Mean	Standard Deviation	Minimum	Q_1	Median	Q_3	Maximum
$174.23	$72.56	$94.75	$124.56	$146.50	$168.50	$412.00

After doing the calculations, he noticed that the maximum monthly bill for $412 should have been $212. How will making the correction affect each of the descriptive statistics below? State *increases, decreases*, or *stays the same*.

   a. mean
   b. standard deviation
   c. median
   d. IQR
   e. range

7. The table shows the descriptive statistics for Ms. Baxter's 2nd period history exam. Terry made the lowest score of 64.

2nd Period History Exam Scores

mean	standard deviation	Minimum	Q_1	Median	Q_3	Maximum
82.2	22.4	64	75	84	88	96

It was later discovered that Terry had cheated, so Ms. Baxter changed his score to 0. What effect does changing Terry's score from 64 to 0 have on each of the descriptive statistics below? State *increases, decreases,* or *stays the same.*

a. mean
b. standard deviation
c. median
d. IQR
e. range

8. A sample of smart phones sold by WorstBuy was collected. Some summary statistics about the prices of the phones (in dollars) are shown below.

Phones Sold at Worst Buy

Mean	Standard Deviation	Median	Q_3	Range
$251.69	$132.42	$246.00	$377.00	$441.00

WorstBuy decides to run a sale on all its smartphones by marking each phone down by $40.00. Find the values of the statistics once the sale price has taken effect.

a. mean
b. standard deviation
c. median
d. Q_3
e. range

Answers

You Try It $2.4.A$

mean = 45.2 cm, standard deviation = 6.5 cm, median = 43.2 cm, and IQR = 15 cm.

You Try It $2.4.B$

mean = 130, standard deviation = 25.

You Try It $2.4.C$

a. increases
b. remains the same
c. decreases
d. remains the same
e. remains the same
f. decreases

This page titled 2.4: Changes to a Data Set is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

3: VISUALIZATIONS OF DATA

3.1: Graphs of Categorical Data
3.2: Histograms, Frequency Polygons, and Ogives
3.3: Other Quantitative Graphs and Data Plots
3.4: Box-and-Whisker Plots
3.5: Normal Distribution and Its Empirical Rule
3.6: Applying a Normal Distribution

This page titled 3: Visualizations of Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

3.1: Graphs of Categorical Data

Learning Objectives

- Organize categorical data in a frequency table.
- Construct graphs of categorical data including bar graphs, circle graphs, and pictographs.
- Interpret categorical data displays.
- Explain the area principle in making and interpreting graphs.
- Apply the three levels of graph comprehension to graphs in various forms.

Introduction

In earlier sections, we explored numerical methods for describing data, such as the mean, median, standard deviation, and the five-number summary. While these measures are essential, there are many situations where we need a more intuitive way to communicate what a data set shows. Graphs and charts allow us to "tell the story" of the data visually.

There are numerous ways to display information graphically, and students encounter many of them throughout the elementary grades and middle school grades. Much of early statistics education focuses on creating, reading, and interpreting these visual representations. This chapter examines several common graphical displays for both categorical and quantitative data, highlighting examples found not only in school curricula but also in media, academic publications, and everyday life.

Frequency Tables

Categorical data consist of information that allows us to place each observation into one of several distinct, non-overlapping categories. A natural first step in organizing such data is to construct a **frequency table**. A frequency table typically begins with two columns: one listing the categories and the other listing the **frequency**, or count, of observations falling into each category.

Example 3.1.1

An insurance company evaluates several risk factors when determining vehicle insurance premiums. One factor they investigate is car color, based on the belief that certain colors may be involved in accidents more frequently. To explore this possibility, the company reviews police reports for recent total-loss collisions. The results are summarized in the frequency table below.

Total Loss Collisions



From this table, *green* appears most often among the collision reports, making *green* the mode. Because these data are categorical nominal, the order of the categories in the table has no inherent meaning. There is no median, since a median requires an ordered scale, and there is no mean, because the data are not numerical quantities that can be added or averaged.

A third column can be added to show relative frequencies, which represent the proportion or percentage of the cars having each color. The total number of collisions in the table is $n = 25 + 52 + 41 + 36 + 39 + 23 = 216$. Each color's percentage is found by dividing its frequency by $n = 216$ and converting the result to a percentage, rounded to the whole percent.

Total Loss Collisions



The exact relative frequencies should sum to be 100% but may be slightly above or below 100% if their values have been rounded as they have here.

Bar Graphs

A **bar graph** is one of the most common and effective ways to display categorical data. In a bar graph, each category is represented by a bar, and the height (in a vertical bar graph) or length (in a horizontal bar graph) of the bar corresponds to the frequency or relative frequency of that category. Bar graphs allow readers to compare categories quickly and visually, making them especially useful in both elementary statistics and real-world data reporting.

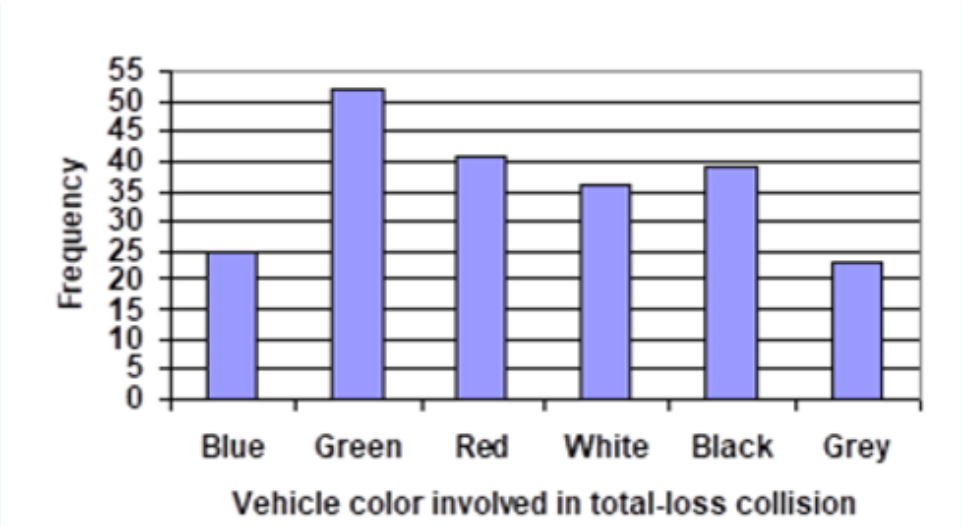
To construct a vertical bar graph, we draw two perpendicular axes.

- The vertical axis typically shows a numerical scale for frequencies or relative frequencies.
- The horizontal axis lists the categories being measured. Because the data are categorical, the horizontal axis does not require a numerical scale.

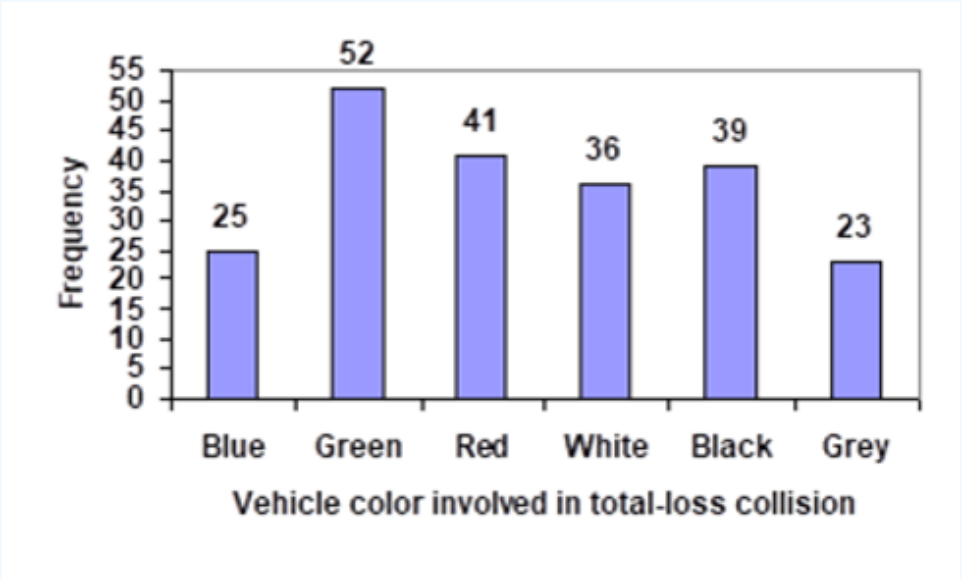
The construction of a bar graph is best illustrated through examples.

Example 3.1.2

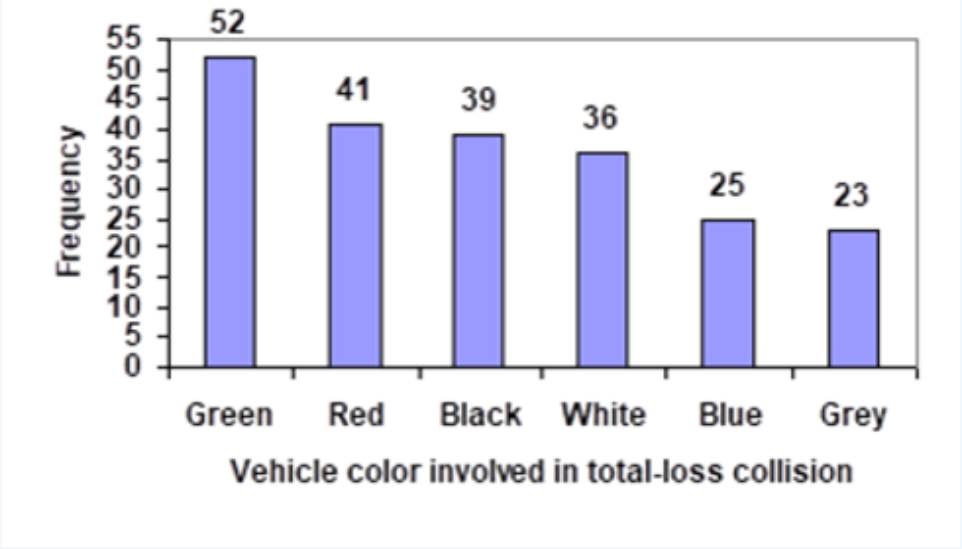
Using the car-color collision data from Example 3.1.1, note that the largest frequency is 52. Therefore, the vertical axis should extend at least to 52. It is often helpful to round slightly upward—such as using 55—so that tick marks can be placed at convenient intervals (for example, every 5 units).



Each bar's height corresponds to the frequency of that color. Horizontal grid lines can make the graph easier to read, but they are optional. Graph paper or technology can simplify construction. Instead of grid lines, some graphs display the frequency values at the top of each bar.



Sometimes it is helpful to reorder the categories from largest to smallest frequency. This arrangement makes comparisons easier and highlights the most common categories. When the bars are arranged in decreasing order of frequency, the graph is called a **Pareto chart**.



A key principle in graphical representation is the **area principle**: the area occupied by each category in a graph should be proportional to the magnitude of the value it represents. In a bar graph, this means that bars must have equal widths, so that differences in bar height accurately reflect differences in frequency. If one category has twice the frequency of another, its bar should be twice as tall—not wider or differently shaped.

Example 3.1.3

A survey asked 30 participants to report the make of their vehicle. The responses were:

F M M M V M F M F V H H F V F H H F M M V H M V V F V H M F

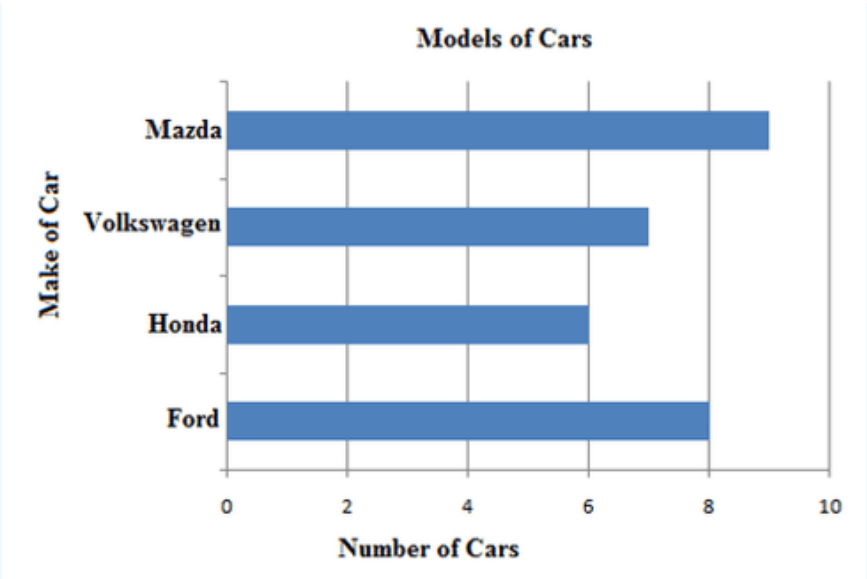
(F = Ford, H = Honda, V = Volkswagen, M = Mazda)

To organize the data, tally marks are useful. The resulting frequency distribution table is

Frequency Table for Car Sample

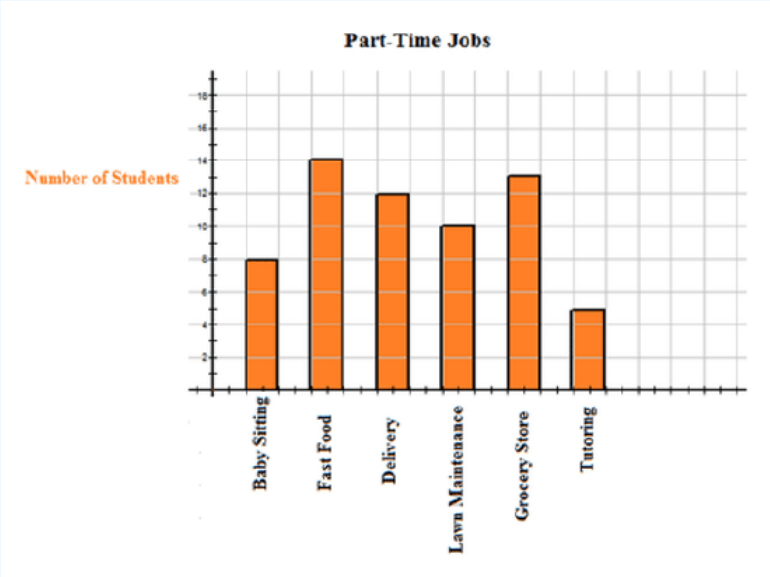
[redacted]

Bar graphs can also be drawn horizontally, with categories listed along the vertical axis and frequencies along the horizontal axis. This orientation is especially helpful when category names are long or when many categories are present.



Example 3.1.4

The bar graph below shows the part-time jobs held by a group of 10th grade students.



a. What is the mode part-time job?

Solution

The mode is *fast food*, since its bar is the tallest.

b. Which part-time jobs employ 10 or more students?

Solution

Fast food, delivery, lawn maintenance, and *grocery store*. All four categories have bars that are 10 units tall or taller.

c. How many more students work in fast food than in tutoring?

Solution

$14 - 5 = 9$ more students

d. What percentage of these students work as delivery drivers?

Solution

First find the total number of students: $n = 8 + 14 + 12 + 10 + 13 + 5 = 62$.

Then, compute the relative frequency: $\frac{12}{62} \approx 19.4\%$.

Side-by-Side Bar Graphs

A **side-by-side bar graph** (sometimes called a **double bar graph**) is used to display and compare two or more related sets of categorical data on the same set of axes. This type of graph is especially helpful when the categories are the same across groups, and the goal is to compare how each group behaves within those categories.

For example, suppose you want to show how many of each type of vegetable you harvested from your garden in June, July, and August. A side-by-side bar graph is an ideal choice. For each vegetable --say, green peppers -- you would place three bars next to each other: one for June, one for July, and one for August. Then you would repeat this pattern for cucumbers, tomatoes, and any other vegetables in your garden.

This arrangement allows you to make two kinds of comparisons at once. First, you can compare the harvest of the *same* vegetable from month to month (for example, whether tomato harvest increased or decreased over the summer). Second, you can compare the *overall* harvest of one vegetable to another (such as whether you harvested more cucumbers than green peppers). By grouping bars within each category and using consistent bar widths, the graph follows the area principle, ensuring that differences in bar height accurately reflect differences in quantity.

A side-by-side bar graph is constructed similarly to a single bar graph, with one important difference: each category contains multiple bars instead of one. The basic steps are:

1. Draw the axes. The horizontal axis lists the categories being compared. The vertical axis shows the frequencies or relative frequencies.
2. Choose an appropriate scale. The scale on the vertical axis must accommodate the largest frequency across all categories.
3. Draw the bars. For each category, draw one bar for each data set. Bars within a category should be grouped together and have equal width.
4. Use color or shading to distinguish groups. Label the axes clearly and include a key (legend) so readers can identify which bar corresponds to which data set.

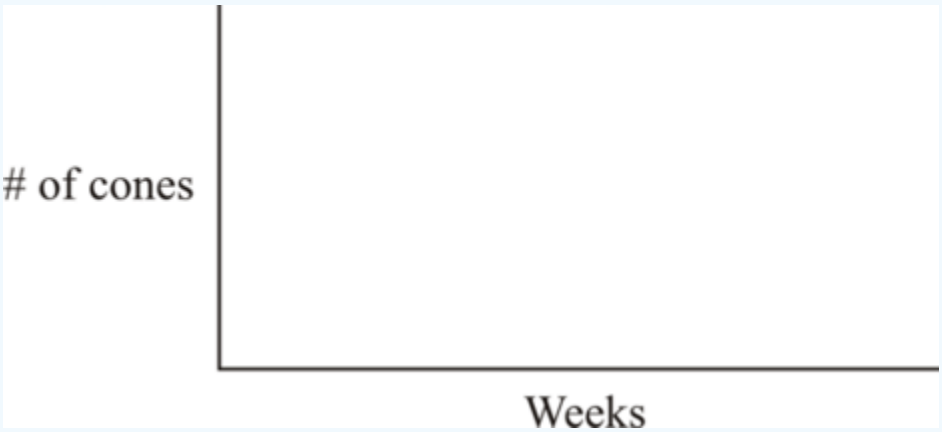
Example 3.1.5

The table below shows the number of ice cream cones sold each week at an ice cream stand during July and August.

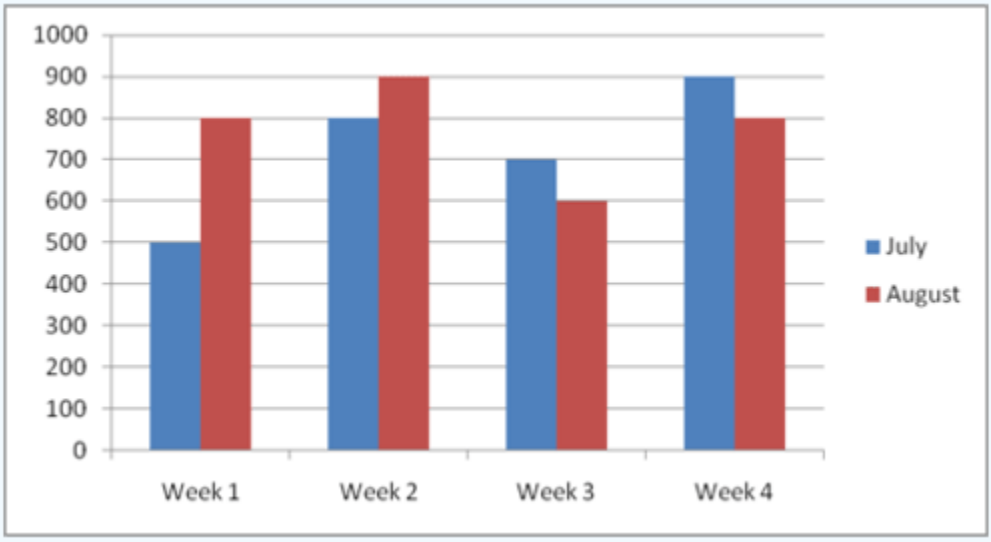
Ice Cream Cones Sold by Month and Week



Because there are two related sets of data -- weekly sales in July and weekly sales in August -- a side-by-side bar graph is appropriate. First, draw the axes. The horizontal axis lists the week numbers, and the vertical axis shows the number of cones sold.



Since the sales values are in the hundreds, a scale from 0 to 1000, counting by hundreds, is reasonable. Next, draw the bars. In this example, blue represents July and red represents August.



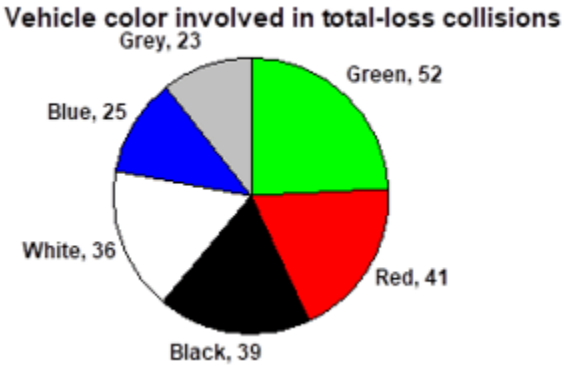
This side-by-side bar graph allows us to compare weekly sales in July and August on a single display. Instead of creating two separate graphs -- one for each month -- we can easily see patterns, similarities, and differences by placing the bars next to each other.

Of course, we could also group the bars differently depending on what we want to emphasize. For example, the graph could have two groups -- one for July and one for August, each showing Week 1, Week 2, Week 3, and Week 4 sales.

Circle Graphs

Circle graphs, also known as **pie charts**, are widely used in newspapers, magazines, business reports, and educational materials. A circle graph visually emphasizes how each category relates to the whole by representing the data as slices of a circle. Each slice corresponds to a category, and the size of the slice reflects that category's relative frequency -- its proportion of the total.

Circle graphs are especially effective when the goal is to show how categories contribute to a whole. They provide an immediate visual impression of proportional relationships, making them a powerful tool for communicating data in both academic and real-world contexts. Here is a circle graph showing the vehicle-color data from the total-loss collisions in Example 3.1.1:



Here we look at two common ways to create a circle graph in the elementary and middle grades:

- Hundredths disk method: The circle is divided into 100 equal wedges, each representing 1% of the whole. This method is intuitive because it directly connects to percentages.
- Angle (degree) method: The circle is divided into 360 degrees. Each category's slice is determined by converting its relative frequency into an angle measure. This method is useful when drawing circle graphs with a protractor.

Both approaches produce the same final graph. Regardless of the method, each slice should be clearly distinguished using category labels and percentages.

The next two examples illustrate both construction methods using the same data set.

Example 3.1.6

The Red Cross Blood Donor Clinic recorded blood types from donors during a successful morning session. The data are shown below:

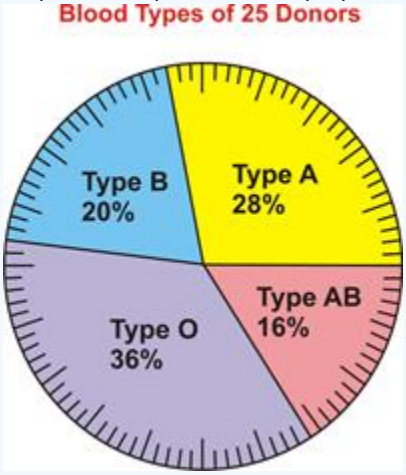
[redacted]

Construct a circle graph to represent the data using a hundredths disk.

Solution

Step 1: Determine the total number of donors: $n = 7 + 5 + 9 + 4 = 25$.

Step 2: Express the proportion of each blood type donation as a relative frequency, in percent:



$\text{Percent} = \frac{f}{n} \cdot 100\%$.

$$\frac{7}{25} \cdot 100\% = 28\%$$

$$\frac{5}{25} \cdot 100\% = 20\%$$

$$\frac{9}{25} \cdot 100\% = 36\%$$

$$\frac{4}{25} \cdot 100\% = 16\%$$

Step 3: Count out the percents (hundredths) for each blood type on the disk.

Step 4: Label each category and relative frequency as percentage.

Example 3.1.7

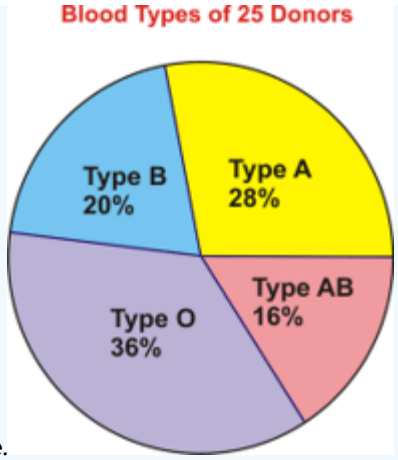
The Red Cross Blood Donor Clinic recorded blood types from donors during a successful morning session. The data are shown below:

[redacted]

Construct a circle graph to represent the data using angles measures.

Solution

Step 1: Determine the total number of donors: $n = 7 + 5 + 9 + 4 = 25$.



Step 2: Express each proportion as its share of the number of degrees in a circle.

$\text{Degrees} = \frac{f}{n} \cdot 360°.$

$\frac{7}{25} \cdot 360° = 100.8°$

$\frac{5}{25} \cdot 1360° = 72°$

$\frac{9}{25} \cdot 360°\backslash{=}129.6°)$

$\frac{4}{25} \cdot 360° = 57.6°$

Step 3: Use a protractor to measure an angle corresponding to each category.

Step 4: Label each category and relative frequency. The angle measures are used to help draw the graph, but they should not be included on the graph

Pictographs

Young children are often introduced to graphs through **pictographs** (or picture graphs). These graphs are highly concrete: instead of bars or numerical scales, they use pictures or icons to represent data. In kindergarten or pre-K, students may even build pictographs on the floor using real objects such as blocks, toys, or classroom supplies. This hands-on approach helps children understand that data can be counted, grouped, and displayed visually.

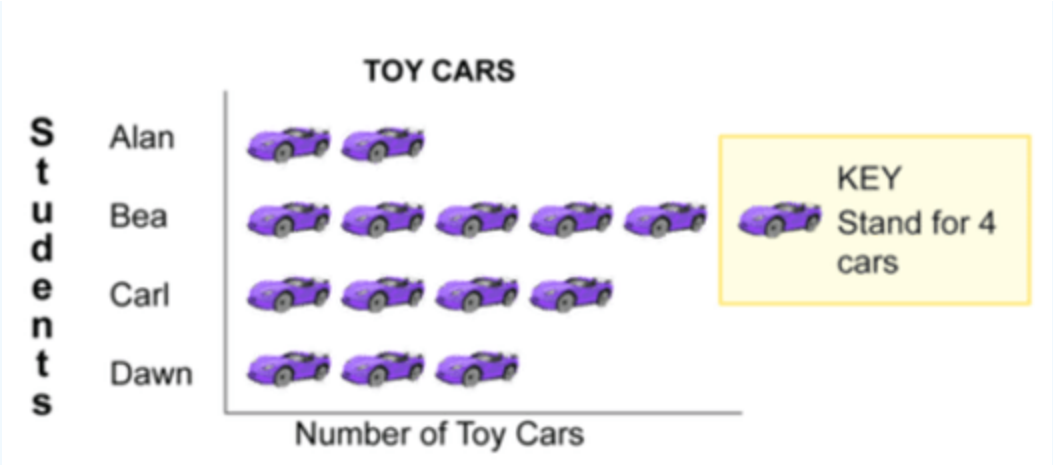
In the simplest pictograph, there is a **one-to-one correspondence** between each picture and the quantity it represents. For example, if a pictograph shows notebook icons, and Greg has 5 notebook icons while Eliza has 2, we interpret this as Greg having 5 notebooks and Eliza having 2 notebooks. Each icon stands for exactly one item. Pictographs form the foundation for more abstract graphs, such as bar graphs, where the length of a bar represents the frequency rather than a count of pictures.

As students begin working with larger numbers, it becomes impractical to draw one picture for every item. At this stage, pictographs evolve: each picture may represent more than one object. This shift requires early algebraic reasoning. Students must understand that a single picture can represent different amounts in different contexts -- just as a variable can represent different values depending on the situation.

An example of a pictograph that does not use a one-to-one correspondence is illustrated in the next example.

Example 3.1.8

Use the pictograph to answer the questions.



a. How many cars does Dawn have?

Solution

There are 3 cars shown in the graph for Dawn. If each icon stands for 4 cars, then Dawn has 3×4 cars.

b. How many more cars does Bea have than Dawn?

Solution

Bea has $5 \times 4 = 20$ cars. We found Dawn has 12 cars in part a. So, Bea has $20 - 12 = 8$ more cars than Dawn.

Pictograms

A **pictogram** is a type of statistical graphic that uses a pictorial representation for data. Unlike bar graphs, which rely on simple geometric shapes, pictograms use meaningful images -- such as dollar signs, cars, or people -- to make a display more visually engaging. Because of their eye-catching design, pictograms appear frequently in newspapers, magazines, advertisements, and online media.

However, pictograms must be interpreted carefully. Their visual appeal can make them misleading if the images are not scaled properly. In a well-constructed pictogram, the pictures themselves must represent the data proportionally. For example, if one data value is double the other, the height of the picture representing the first quantity should be twice as tall as the second picture while their widths remain constant. When both height and width change, the area of the image changes disproportionately, violating the *area principle*: the area used to represent a value must be proportional to the magnitude of that value.

If this principle is ignored, the graph can exaggerate differences and distort the message as this next example illustrates.

Example 3.1.9

A labor union wants to stress that managers earn twice the salaries of workers. To convey this idea, a graphics designer makes a money bag "twice as big" for managers than for workers as shown below. Critique this graph.



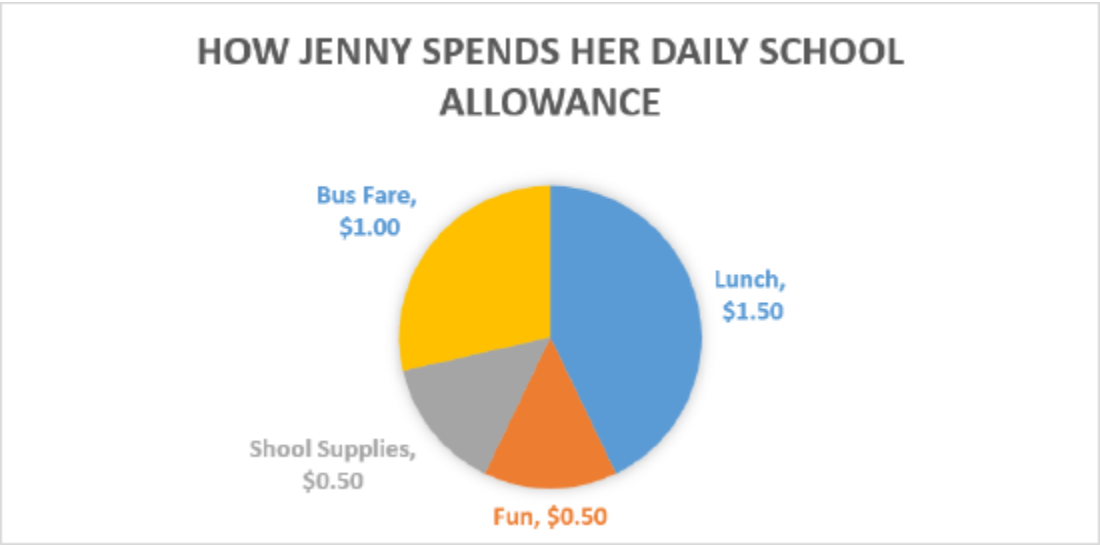
Solution

The graphic is misleading. At first glance, the manager's bag appears far larger than twice as big -- closer to four times the size of the worker's bag. This is because the designer increased both the height and the width of the picture. When a picture is twice as tall and twice as wide, its area becomes four times as large.

This creates a misleading impression and violates the area principle. The graph visually suggests a much larger salary gap than the data support.

Levels of Graph Comprehension

Graphs allow students not only to read information but also to interpret, analyze, and extend it. While elementary and middle school learners must first be able to read data directly from a graph, deeper understanding emerges when they can make comparisons, draw inferences, and apply information beyond what is explicitly shown. Consistent with the reading comprehension literature, graph comprehension can be described in three distinct levels: *reading the data, reading between the data*, and *reading beyond the data*. These levels apply to all graph types. Examples of questions written at each of these three levels are illustrated below with this circle graph:



Reading the Data

At this level, students engage in a literal reading of the graph. They extract facts that are explicitly stated, relying on labels, titles, and directly presented quantities. This is a low-level cognitive task requiring no interpretation or inference.

Examples of reading-the-data questions:

- How much does Jenny spend on bus fare each school day?
  - *Answer: $1.00*
- Whose school day allowance does this graph show?
  - *Answer: Jenny*

Reading Between the Data

This level requires students to interpret and integrate information from the graph. Rather than simply lifting facts, readers must compare quantities, identify relationships, or perform mathematical operations such as addition, subtraction, multiplication, or division. Standardized assessments frequently target this level because it requires at least one step of logical reasoning.

Examples of reading-between-the-data questions:

- How much more does Jenny spend on lunch than on bus fare?
  - *Answer: $0.50. This comparison requires subtraction.*
- On which expense does Jenny spend the most money?
  - *Answer: Lunch. Although the fact appears obvious, determining it requires comparing the sizes of the slices or the amounts.*
- How much is Jenny's total school day allowance?
  - *Answer: $3.50. This requires adding all the expense amounts.*

Reading Beyond the Data

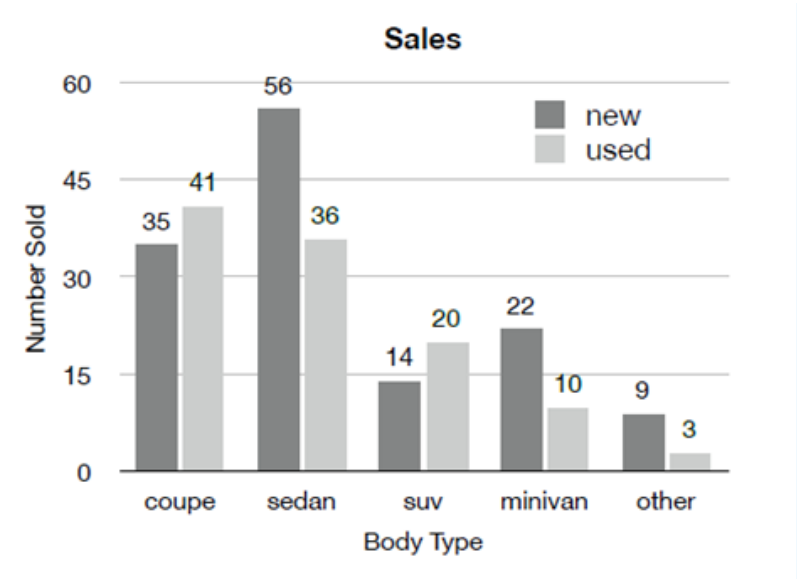
At this level, students must use information not provided in the graph. They draw on background knowledge, make predictions, or apply external facts. Whereas *reading between the data* involves inferences based solely on the graph, *reading beyond the data* requires inferences supported by the reader's own knowledge base or external information.

Examples of reading-beyond-the-data questions:

- How much money does Jenny spend on lunch during a full school week?
  - *Answer: $7.50. This requires multiplication and outside knowledge that a school week has five days.*
- Jenny plans to save the "fun" portion of her allowance to buy a movie ticket costing $8.00. How many days will she need to save?
  - *Answer: 16 days. This requires division and outside information—the $8.00 ticket price.*

Example 3.1.10

The side-by-side bar graph shows last year's sales of autos at Carl's Cars by body style and condition (new/used). Use this graph to answer the following questions and tell the level of graph comprehension the question targets.



a. How many used minivans were sold?

Solution

10 minivans. This fact is read directly from the graph so this is at the level of *reading the data*.

b. For which body type(s) is “buying used” more popular than “buying new”?

Solution

Coupe and SUV. To answer this question, we compare the number of new vehicles to the number of used vehicles for each of the 5 body types. This is a *reading between the data* problem.

c. Do you think most of Carl’s customers tend to have children? Why or why not?

Solution

Probably not. Fewer family cars (like mini-vans and SUVs) were sold relative to smaller cars (like coupes and sedans.) Families with children are more likely to purchase larger type vehicles. This is an example of *reading beyond the data* type question. You needed some background information about what families might prefer and also needed to know which types of vehicles are small and large.

d. What percentage of the cars sold at Carl’s Cars were coupes?

Solution

About 31%. 76 coupes were sold out of a total 246 vehicles. This is a *reading between the data* problem. All the information needed to answer the question is in the graph. To find the solution, operations such as addition and division were used.

Section Summary

A frequency table for a categorical variable lists each category along with its frequency (count) and relative frequency (percentage). Three common graph types used to display categorical data are the bar graph, circle graph, and pictograph.

Circle graphs represent the distribution of a categorical variable as a “pie,” with each slice sized according to the category’s relative frequency. Although they can be cumbersome to construct by hand, one can think of the circle as consisting of 100 percent or 360 degrees and assign portions accordingly. A circle graph must include all categories in the data set and is most

effective when the goal is to highlight each category's relationship to the whole.

Bar graphs display each category as a bar and can represent either frequency or relative frequency. They are generally more flexible than circle graphs because they allow comparisons across categories whether or not the entire data set is shown. Bar graphs are especially useful when comparing quantities measured in the same units.

Pictographs, often introduced in the primary grades, use pictures or symbols to represent counts within categories. When the symbols do not correspond one-to-one with the data, a clear key is required.

Across all graph types, it is essential to follow the area principle: the area occupied by a graphical element should be proportional to the magnitude of the data it represents.

Research identifies three levels of graph comprehension: reading the data, reading between the data, and reading beyond the data. Reading the data involves extracting literal information directly from the graph. Reading between the data requires manipulating or comparing the presented values. Reading beyond the data requires using information not shown in the graph—such as background knowledge or external facts—to answer a question.

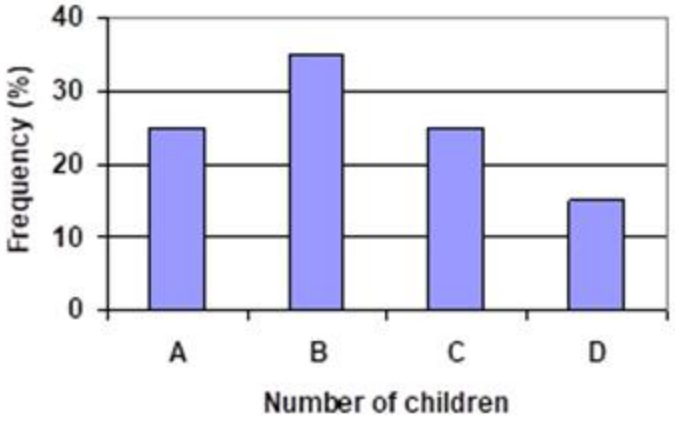
Section 3.1 Review Questions

1. Jasmine is a customer service representative for a cable company. At the end of every call, customers are invited to rate their experience with Jasmine's service on the call. The bar graph below shows the data for one shift.



   a. At which level of measurement is the data for '*customer service rating*'?
   b. How many ratings did Jasmine receive on her shift?
   c. What percentage customers were satisfied?
   d. How many more customers were extremely satisfied than extremely dissatisfied?

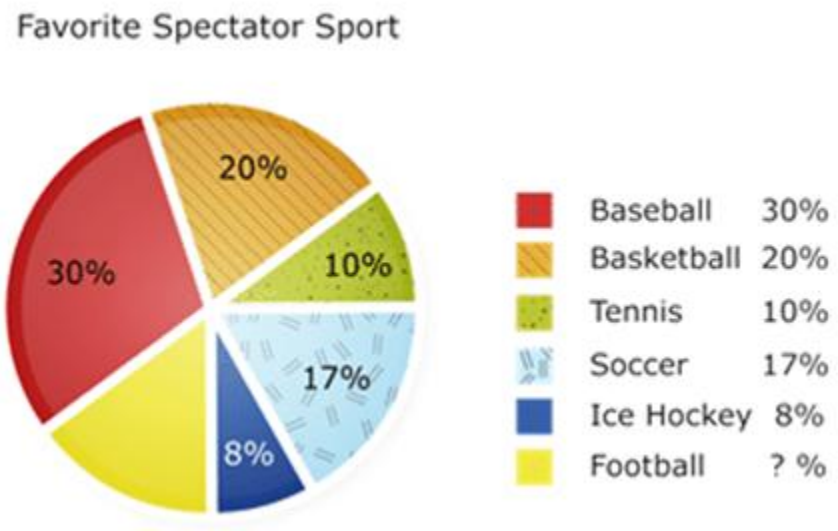
2. The bar graph shows the percentage of students who received each letter grade on their last English paper. The class contains 20 students. What number of students earned an A on their paper?



3. The table shows the number of students in the 7th grade who are studying each foreign language. Assume each student studies only one language. Complete the table so that it shows the relative frequency and the corresponding measure of the central angle used for each category in a circle graph.



4. The graph shows results of a survey of 300 people about their favorite spectator sport.

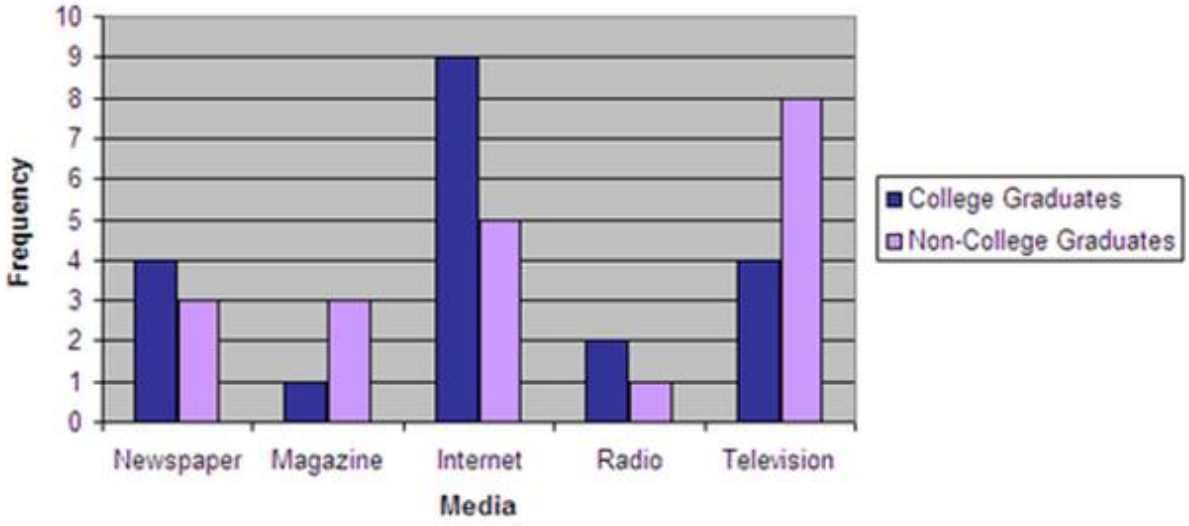


   a. What percent of the people surveyed said that football was their favorite sport?
   b. How many people said that basketball was their favorite sport?
   c. How many more people said that soccer was their favorite sport than ice hockey?
   d. Convert the data shown by the circle graph to a Pareto graph that shows frequency of responses along the vertical axis.

5. The table shows how much money the students in the seventh grade have raised so far for a class trip. Make a well-labeled bar graph and a circle graph that shows the data.

[redacted]

6. Some college graduates and some non-college graduates were asked their primary way of following the news, and the results are shown in the side-by-side bar graph below. Use the side-by-side graph to answer the following questions.



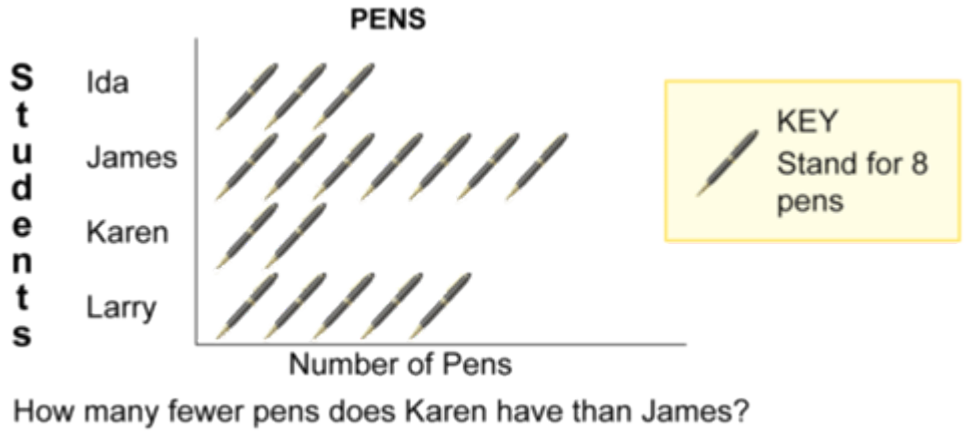
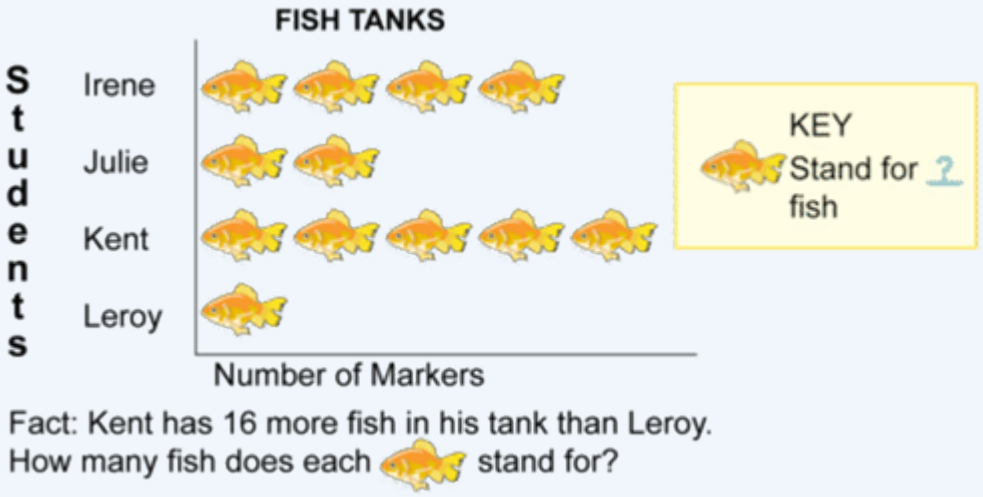
   a. How many more non-college graduates use television for news than the newspaper?
   b. When college graduates and non-college graduates are combined, which form of the media is used the least to follow the news?
   c. What percentage of the college graduates use the internet to follow the news?

7. The guidance counselor at USA High School wanted to know what future plans the graduating class had. She took a survey to determine the intended plans for both boys and girls in the school's graduating class. Use the data to draw a side-by-side bar graph to represent the data. Draw any conclusions that you can from the resulting chart.

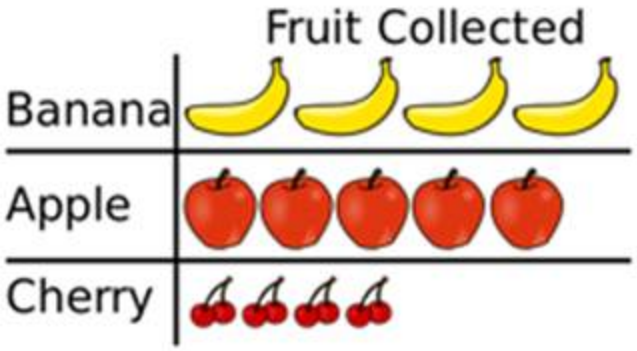
[redacted]

8. The following problems are from a 3rd grade curriculum. Answer each question.

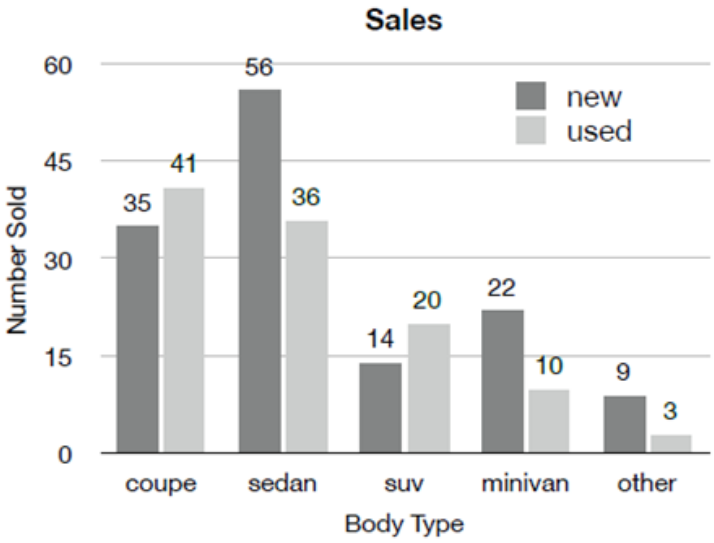
Problem 8

a.	 How many fewer pens does Karen have than James?
b.	 Fact: Kent has 16 more fish in his tank than Leroy. How many fish does each fish icon stand for?

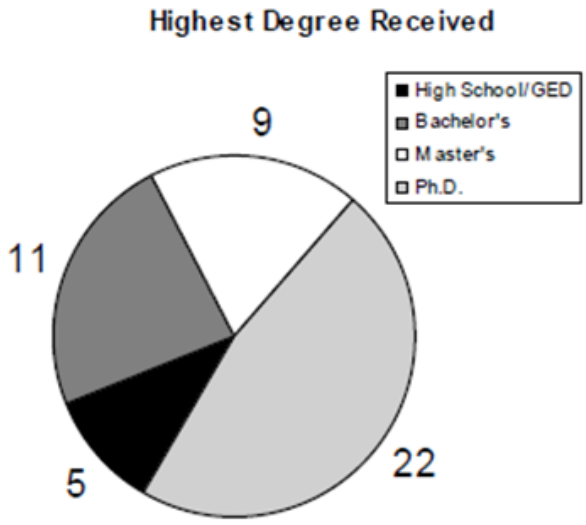
9. Jacob asked his classmates to donate fruit to make fruit baskets for their teacher. He made a graph of the fruit he collected. Describe why this graph may be misleading and give a remedy for correcting it.



10. The side-by-side bar graph shows last year's sales of autos at Carl's Cars by body style and condition (new/used). Answer the questions *and* tell the level of graph comprehension each question addresses.



a. How many new cars were sold at Carl's Cars last year?
b. What was the weekly average number of cars sold at Carl's Cars last year?
c. How many used sedans were sold at Carl's Cars last year?
d. What percentage of the sedans sold at Carl's Cars were used?
e. How many more new sedans than used sedans were sold at Carl's Cars?
f. If Carl's Cars makes an average profit of $500 on each used minivan, what was the profit last year on used minivans?



11. A survey was given to several employees. Among the survey questions was one that asked the highest educational degree attained by the employee. The data obtained from that question are displayed in the circle graph shown here.
    a. Write and answer your own *reading the data* question about this graph.
    b. Write and answer your own *reading between the data* question about this graph.
    c. Write and answer your own *reading beyond the data* question about this graph.

This page titled 3.1: Graphs of Categorical Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

3.2: Histograms, Frequency Polygons, and Ogives

Learning Objectives

- Organize quantitative data in grouped and ungrouped frequency tables.
- Construct a histogram, frequency polygon, and ogive.
- Interpret information from histograms, frequency polygons, and ogives.
- Explain implications of histogram shape on measures of center and spread.
- Analyze a histogram to estimate center and spread.

Introduction

As students begin working with larger sets of quantitative data, organizing those values becomes essential for revealing patterns and meaningful structure. Numerical summaries such as the mean, median, and standard deviation are informative, but they cannot fully show how the data are distributed. Visual displays complement these summaries by providing an intuitive picture of the distribution's behavior. In this section, we learn to construct and interpret frequency tables and foundational graphs for quantitative data -- histograms, frequency polygons, and ogives -- tools that help us compare distributions, recognize shapes such as symmetry and skewness, and estimate important measures of center and spread.

Ungrouped quantitative data list each numerical value exactly as recorded, preserving full detail and allowing precise descriptive statistics. **Grouped data**, on the other hand, place values into class intervals and record only the frequencies within those intervals. This approach highlights broader patterns and makes large data sets easier to summarize, though it no longer retains the exact individual values. In this section we explore organizing, displaying, and describing both ungrouped and grouped data using frequency tables, histograms, frequency polygons, and ogives.

Working with Ungrouped Quantitative Data

Frequency Table

When working with ungrouped quantitative data, the first step is to organize the raw values so the overall distribution becomes easier to understand. An **ungrouped frequency table** provides this initial structure by listing each distinct data value and the number of times it occurs. This simple organization allows us to see which values are most common, how the data are spread out, and whether any observations stand out as unusually high or low. Tally marks are often helpful during construction because they provide an accurate running count as each observation is recorded.

Histogram

Once the data have been organized into an ungrouped frequency table, we can translate that information into a **histogram**. A histogram is a graphical display for quantitative data in which each bar represents a *bin* (a single value or interval of values) and the height of the bar shows the frequency. For ungrouped data, each individual value becomes its own bin, so the bars sit directly next to one another without gaps. This adjacency reflects the continuous nature of the number line and distinguishes histograms from bar graphs used for categorical data. Histograms allow us to see the shape of a distribution -- its peaks, clusters, gaps, and overall spread -- at a glance.

Example 3.2.1

Each student in a class reports how many plastic beverage bottles they use in a week.

$$\begin{array}{cccccccccccccc} 6 & 4 & 7 & 7 & 8 & 5 & 3 & 6 & 5 & 7 & 6 & 7 & 6 & 6 & 7 \\ 7 & 5 & 2 & 6 & 1 & 3 & 5 & 4 & 7 & 4 & 5 & 4 & 6 & 5 & 3 \end{array}$$

The variable “number of plastic bottles used” is quantitative and discrete. Because the responses fall within a relatively short range of whole numbers, each value can serve as its own bin.

After tallying the data, we have this ungrouped frequency table listing each value from 1 through 8 and the number of students who reported that value.

Frequency Table for Water Bottle Usage

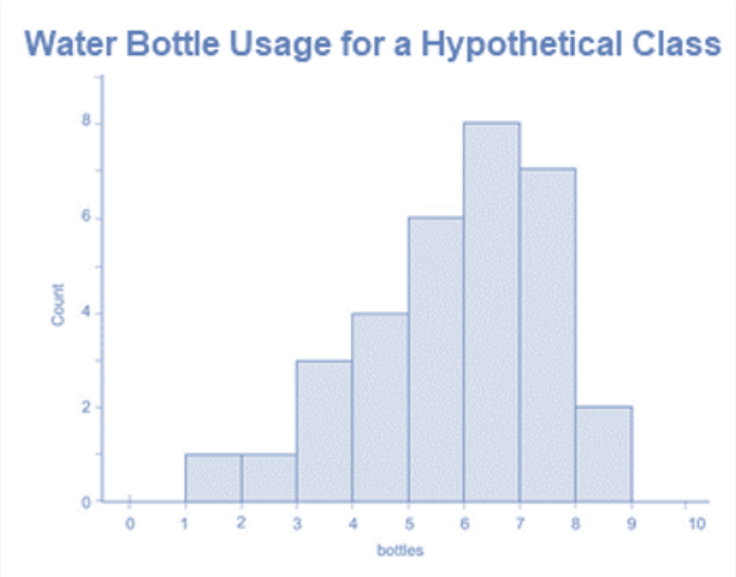
Number of Water Bottles Used per Week	Tally	Frequency
1	|	1
2	|	1
3	| | |	3
4	| | | |	4
5	𝍸 |	6
6	𝍸 | | |	8
7	𝍸 | |	7
8	| |	2

To construct the histogram, place the variable “number of water bottles used” on the horizontal axis and the frequency on the vertical axis. The horizontal axis is divided into equal-width intervals called bins. In this example, each data value forms its own class, so we use a bin width of 1. You can think of each bin as a container that gathers all data values falling within its interval.

- First bin — This bin begins at 1 and extends to 2. It contains the one student who used exactly 1 bottle.
- Second bin — This bin begins at 2 and extends to 3. It contains the one student who used exactly 2 bottles.
- Third bin — This bin begins at 3 and extends to 4. It contains the three students who used exactly 3 bottles.
- Continuing bins — This pattern continues until the final bin, which begins at 8 and extends to 9 and contains the two students who used exactly 8 bottles.

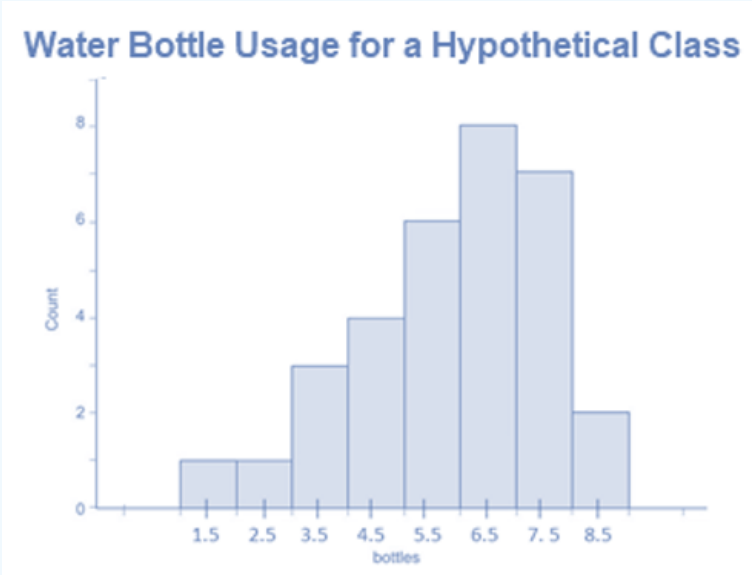
The height of each bar represents the number of students who used that particular number of bottles. From the frequency table, the tallest bar corresponds to the bin from 6 to 7, with a height of 8.

Here is the completed histogram:

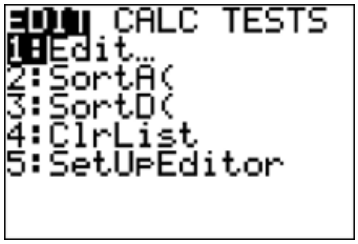
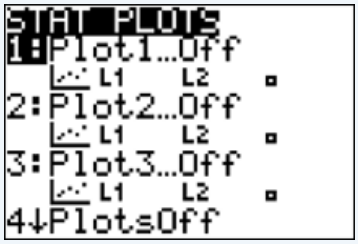
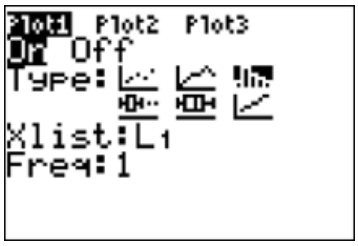
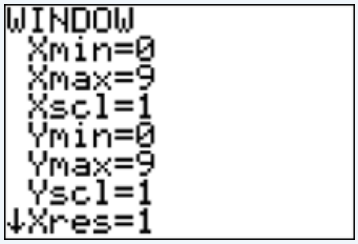


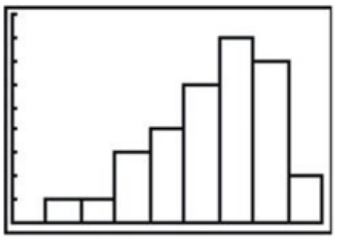
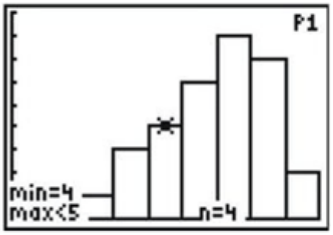
It is important when drawing a histogram to ensure that the bins fully cover the range of the data. The final bin must extend far enough to include the largest observed value. This often requires stretching the horizontal axis slightly beyond the last data point. In this example, stopping the axis at 8 would incorrectly omit the two students who used 8 bottles, because the value 8 belongs in the bin that begins at 8 and extends to 9. Careful attention to bin boundaries helps prevent misrepresenting the distribution.

In many published histograms -- such as those appearing in newspapers, magazines, or online -- the midpoint of each bin is labeled rather than the interval endpoints. Some graphing software also defaults to midpoint labeling unless the user specifies otherwise. These midpoints are known as *class marks*.



Regardless of the labeling convention, the underlying structure of the histogram remains the same: each bar represents a bin, and its height shows the frequency of observations in that interval.

1.	To draw a histogram, you must first enter the data into a list as we have done in the past. Press [**STAT]** and then choose **1:Edit** to input the data.	
2.	Press **[2ND] [Y=]** to enter the **STAT PLOTS** menu.	
3.	You can plot up to three statistical plots at one time. Choose **Plot 1**. Move the cursor to the word **On** and press **[ENTER]** to turn the plot on. Change the **Type** of plot to a histogram, and choose L1 for **Xlist.** Enter '1' for **Freq** by pressing [**2ND][A-LOCK]** to turn off alpha lock, which is normally on in this menu. If you have already made a frequency table, you enter the values of the variables in L1 and the frequencies in L2 as we did in Chapter 1 rather than typing in all the raw data.	
4.	We need to set a window. Press **[WINDOW]** and enter an appropriate window to display the plot. In this case, **Xscl** is what determines the bin width. Also notice that the **Xmax** value needs to extend right to 9 to show the last bin, even though the data values stop at 8. Enter all the values shown.	

5.	Press **[GRAPH]** to display the histogram. If you press **[TRACE]** and then use the left or right arrows to trace along the graph, notice how the calculator uses the notation to properly represent the frequency of values in each bin. **Note:** Instead of setting the WINDOW manually, you can let the calculator set its own window by pressing **[ZOOM]** and then choosing **9:ZoomStat**.	 

Relative Frequency Histograms

In addition to displaying raw frequencies, it is often useful to show the relative frequency of each bin -- that is, the proportion of the data set that falls within each interval. A **relative frequency histogram** uses the same bins as the original histogram, but the vertical axis represents relative frequency rather than count. Each bar's height is the fraction (or percentage) of observations in that bin.

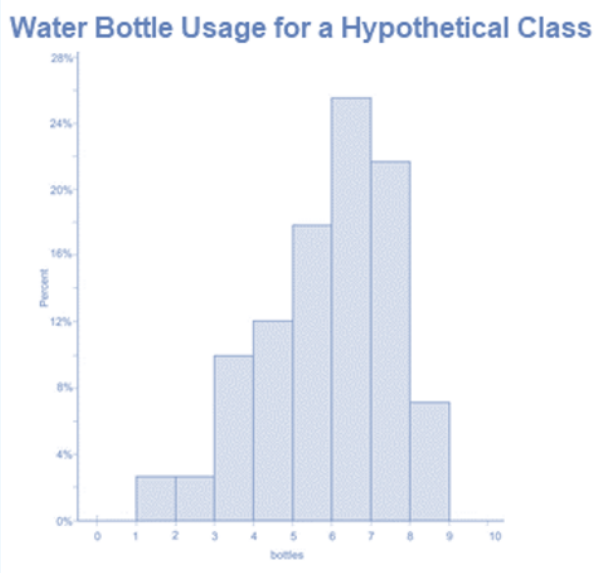
Relative frequency histograms are especially helpful when comparing two data sets of different sizes. Because the vertical axis is scaled by proportion, not by the number of observations, the shapes of the distributions can be compared directly. A reader can quickly see which intervals contain larger or smaller portions of the data.

Example 3.2.2

Show the data in Example 3.2.1 as a relative frequency histogram.

Solution

There is only one data value in the first bin of the histogram. This represents $frac132 \approx 0.33$ of the total data. Thus, the vertical bar for the bin extends upward to 3%. Likewise, the height of the second bar has height 3%. The overall shape of the histogram is the same no matter if you use frequency or relative frequency along the vertical axis. When making a relative frequency histogram, be sure to clearly label the axis as '*percent*.'



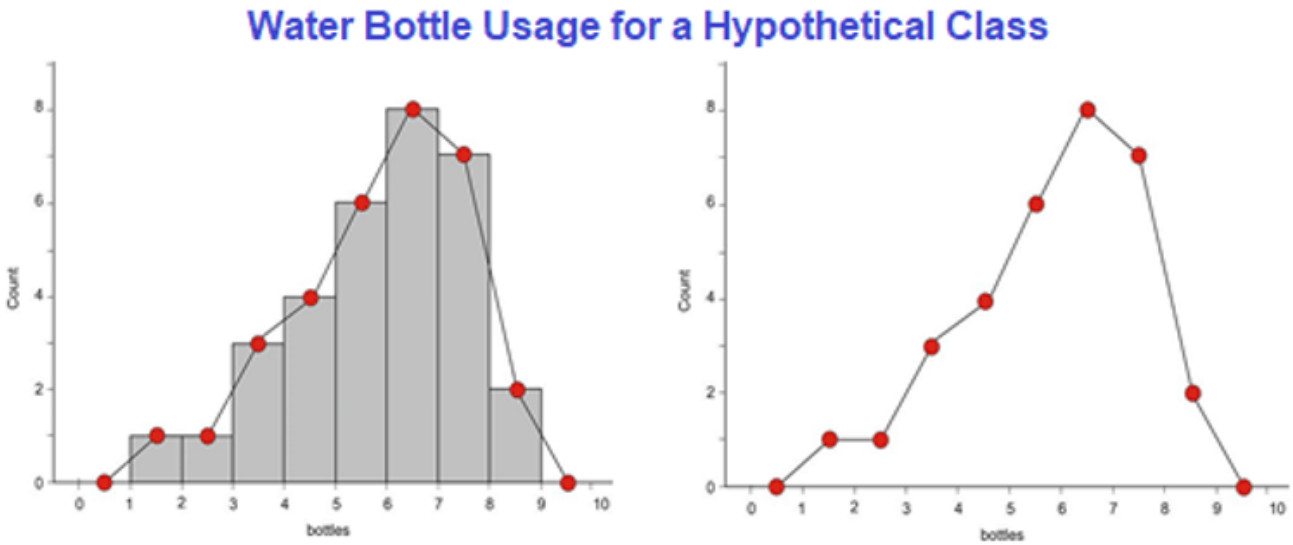
Frequency Polygons

A **frequency polygon** provides another way to visualize the distribution of quantitative data. It is similar to a histogram, but instead of using bars, the frequencies are plotted as points and connected with straight line segments.

To construct a frequency polygon for the students' bottle-use data:

1. Identify the midpoint (class marks) of each class interval.
2. Plot a point at each midpoint at a height equal to the bin's frequency.
3. Connect the points with straight line segments.
4. To close the polygon, plot one additional midpoint below the smallest class and one above the largest class, each with frequency 0, and connect these points to the rest of the graph.

When superimposed on the histogram, the frequency polygon traces the tops of the bars, creating a smooth outline of the distribution. When shown alone, it highlights the overall shape more clearly. Frequency polygons are especially useful when comparing multiple data sets on the same axes because the lines can be overlaid without obscuring one another.



Cumulative Frequency Histograms and Ogives

Often, it is helpful to understand how the data accumulate across the range of values. To do this, we extend the frequency table by adding a **cumulative frequency** column. The cumulative frequency for a bin is the total number of observations in all bins **up to and including** that bin.

For the bottle-use data:

- The cumulative frequency for 1 bottle is 1 student.
- The cumulative frequency for 2 bottles is 2 students $(1+1)$
- The cumulative frequency for 3 bottles is 5 students $(1+1+3)$.
- The cumulative frequency for 4 bottles is 9 students $(1+1+3+4)$.

Continuing in this way produces a complete cumulative frequency table. The cumulative frequency for the final bin must always equal the total number of observations in the data set.

Frequency Table for Class Usage of Water Bottles

Number of Water Bottles Used per Week	Frequency	Cumulative Frequency
1	1	1
2	1	$1+1=2$
3	3	$2+3=5$
4	4	$5+4=9$
5	6	$9+6=15$

Number of Water Bottles Used per Week	Frequency	Cumulative Frequency
6	8	$15 + 8 = 23$
7	7	$23 + 7 = 30$
8	2	$30 + 2 = 32$

A **cumulative frequency histogram** displays these cumulative frequencies using the same bins as the original histogram, but the height of each bar represents the cumulative total rather than the frequency for that bin alone. This type of histogram shows how quickly the data accumulate and highlights where large increases occur.

From the cumulative frequency histogram, we can construct an **ogive** (pronounced *OH-jive*). An ogive is similar to a frequency polygon, but it plots cumulative frequencies instead of bin frequencies. To create an ogive,

1. Plot a point at each upper class boundary at the cumulative frequency.
2. Connect the points with straight line segments.
3. Begin the graph at the lower boundary of the first class with cumulative frequency 0.

Using the data from Example 3.2.1.

- The point plotted at 2 corresponds to 1 student (all students who used one or fewer bottles).
- The point plotted at 3 corresponds to 2 students (all students who used two or fewer bottles).
- The point plotted at 4 corresponds to 5 students (all students who used three or fewer bottles).

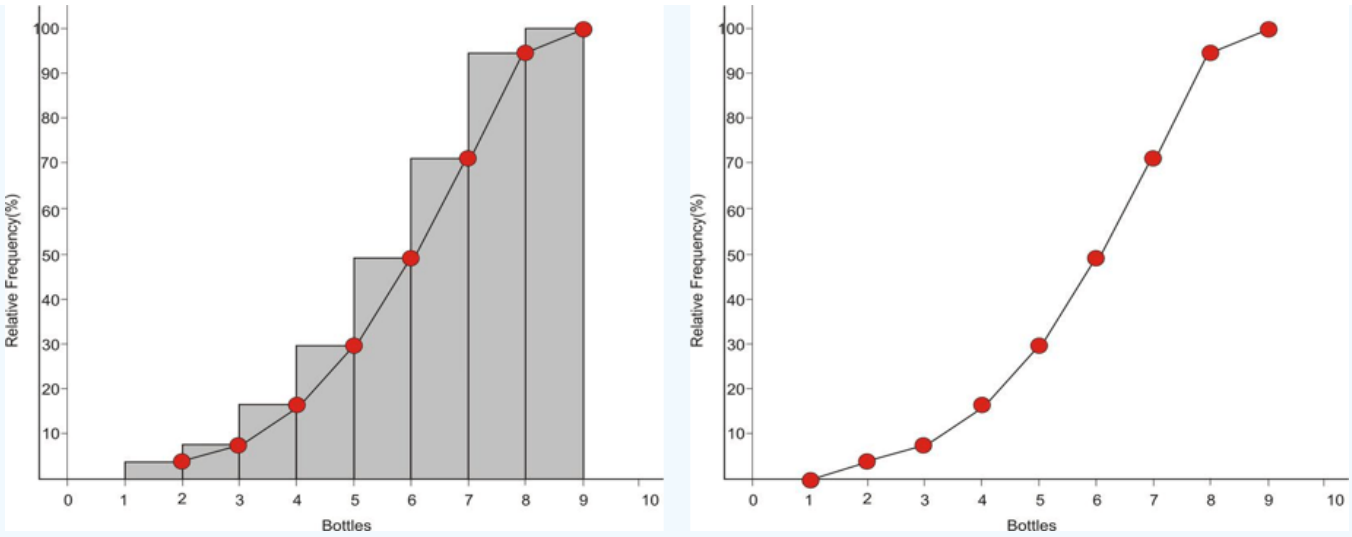
This process continues through the remaining intervals. The ogive rises steadily from left to right, showing how the data accumulate across the distribution.

Example 3.2.3

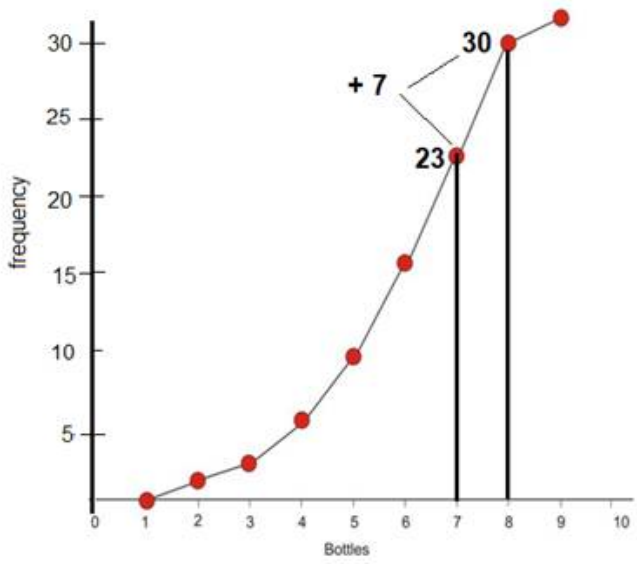
Show the data in Example 3.2.1 as a cumulative frequency histogram and as an ogive.

Solution

The frequency histogram and the ogive are shown here:



An important feature of ogives is that you can recover the frequency of any bin by subtracting cumulative frequencies. For example, if 23 students had accumulated by 7 bottles and 30 students had accumulated by 8 bottles, then the number of students in the interval [7, 8) is $30 - 23 = 7$ students.



Ogives are especially useful for estimating medians, quartiles, and percentiles because they show cumulative behavior directly. You can explore this idea further by examining an ogive or comparing it with a frequency polygon.

Working with Grouped Data

Frequency Table

When quantitative data are continuous or when the values span a wide range, listing every distinct value in an ungrouped frequency table is neither practical nor helpful. In such cases, it is more effective to organize the data into **grouped classes**, or **intervals**, and record how many observations fall within each interval. Grouped frequency tables allow us to summarize large or varied data sets in a way that reveals meaningful patterns in the distribution.

Grouping becomes necessary when the data are continuous, when there are many distinct discrete values, or when the range is large. Instead of letting each value form its own class, we define arbitrary but sensible intervals of equal width to create a clearer

picture of how the data are distributed.

Example 3.2.4

Suppose a homework system reports how long (in minutes) college students spent on a recent homework assignment.

6.2	7.1	7.5	8.0	8.4	9.1	10.3	11.2	12.0
12.6	13.1	14.8	15.0	16.3	17.9	18.4	20.7	23.9

Organize the data into a grouped frequency table and construct a histogram. Use a class width of 4 minutes and begin the first class at 6 minutes.

Solution

Because the values range from about 6 to 24 hours, it would not be useful to create a separate bin for each number. Instead, we define class intervals of width 4 minutes:

- [6, 10)
- [10, 14)
- [14, 18)
- [18, 22)
- [22, 26)

Recall from algebra, a bracket [or] means the endpoint is included. A parenthesis (or) means the endpoint is excluded. In statistics, it is standard practice to include a boundary value in the next interval. For example, [6,10) includes all values from 6 up to, but not including, 10. The value 10 belongs to the next interval, [10,14).

Because the data are already ordered, we can count how many values fall into each interval without using tally marks. The completed grouped frequency table is

Frequency Table for Homework Duration



This table shows a clear unimodal pattern: the highest concentration of study hours falls in the interval [6, 10), with frequencies tapering off to the right.

Histogram

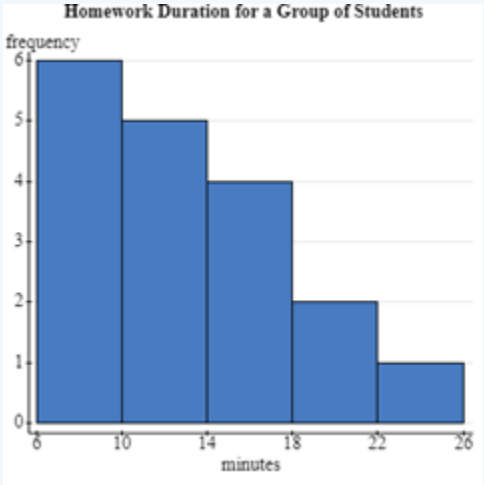
Once the class intervals have been chosen and the frequencies counted, the grouped frequency table becomes the foundation for constructing several graphical displays. Each graph highlights a different aspect of the distribution. A histogram or relative frequency histogram can be used to emphasize the overall shape of the data by showing how values are distributed across the intervals. A frequency polygon provides a connected-line version of the same information, making patterns such as peaks or skewness easier to compare across multiple data sets. If the goal is to show how quickly the data accumulate as we move through the intervals, an ogive offers a clear picture of that progression. Together, these displays allow us to move smoothly from a grouped frequency table to visual representations that reveal structure, trends, and meaningful features of the distribution.

Example 3.2.5

Use the grouped frequency table in the previous example to make a frequency histogram.

Solution

Because the intervals represent continuous data, the bars touch unless a class has frequency zero. In this example, all intervals contain at least one observation, so there are no gaps.



Example 3.2.6

Answer the questions using the frequency table that summarizes the data when several students were asked how many college credits they had completed.



a. What is the total number of data values in this data distribution?

Solution

Adding the frequencies of each class, we have $4 + 7 + 1 + 0 + 3 + 5 = 20$.

b. What class width (bin size) is used to group the data?

Solution

Subtract any two consecutive lower class limits or any two consecutive upper class limits. For example, $16–9 = 7$.

c. What is the modal class?

Solution

The class with the highest frequency is [16,23).

d. If an additional class were added to the end of the table, what would be the upper and lower class limits?

Solution

Add the class width 7 to the last lower and upper class limits to get [51,58).

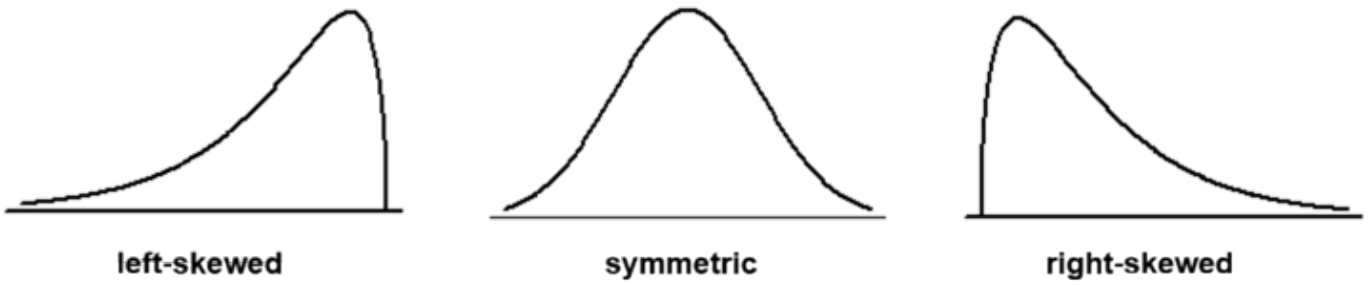
Shape, Center, and Spread in Graphical Displays

When we analyze a frequency table or histogram, we are not simply counting values -- we are trying to understand how the data behave as a whole. The **shape** of the distribution provides the first major clue. Shape tells us where data cluster, how they taper off, and whether unusual features such as gaps or multiple peaks appear. These visual characteristics help us decide which numerical summaries are most appropriate and how to interpret them.

A histogram is especially useful because it shows how frequently values fall into each interval. When the bars rise to form a single, smooth peak, the distribution is **unimodal**. If that single peak is centered and the left and right sides mirror each other, the distribution is **symmetric** and often described as **bell-shaped**. When two distinct peaks appear, the distribution is **bimodal**, often suggesting two different subgroups within the data. Distributions with more than two distinct peaks are referred to as **multimodal.** A nearly flat histogram indicates a **uniform** distribution, where all values occur with roughly equal frequency.

Skewness describes how a distribution deviates from being symmetric. In a symmetric distribution, the left and right sides of the histogram mirror each other. When a distribution is **skewed-right,** the bulk of the data lies on the left with a long tail tapering off to the right. Conversely, a **skewed-left** distribution has most data on the right and a long tail extending left towards the smaller values.

Smoothed versions of histograms called **density curves** allow us to describe shape more precisely. A density curve highlights the overall pattern without being influenced by small sample fluctuations.

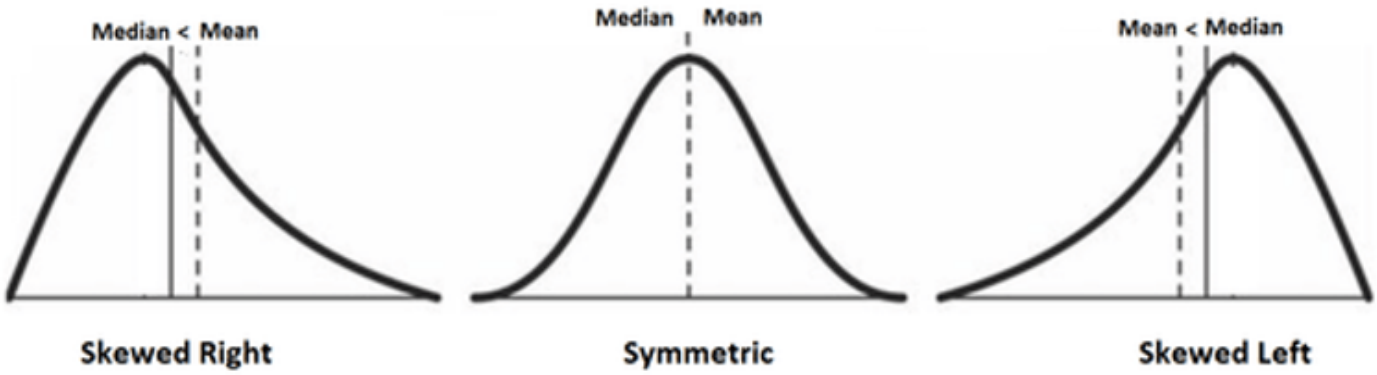


Connecting Shape to Measures of Center

Recall that the center of a distribution is typically described using the mean or median. The numerical relationship and choice between these depends heavily on the shape.

- In symmetric distributions, the mean and median fall at nearly the same location. The balance of the left and right tails ensures that extreme values do not pull the mean away from the center.
- In right-skewed distributions, the long right tail contains larger values that pull the mean to the right. The median, being resistant to extreme values, remains closer to the bulk of the data.
- In left-skewed distributions, the long left tail contains smaller values that pull the mean to the left. Again, the median stays near the center of the main cluster.

This relationship is so consistent that, once we identify the shape, we can often predict the ordering of mean and median before computing them. The images below illustrate the numerical comparisons of the mean versus the median.



Spread and the Influence of Shape

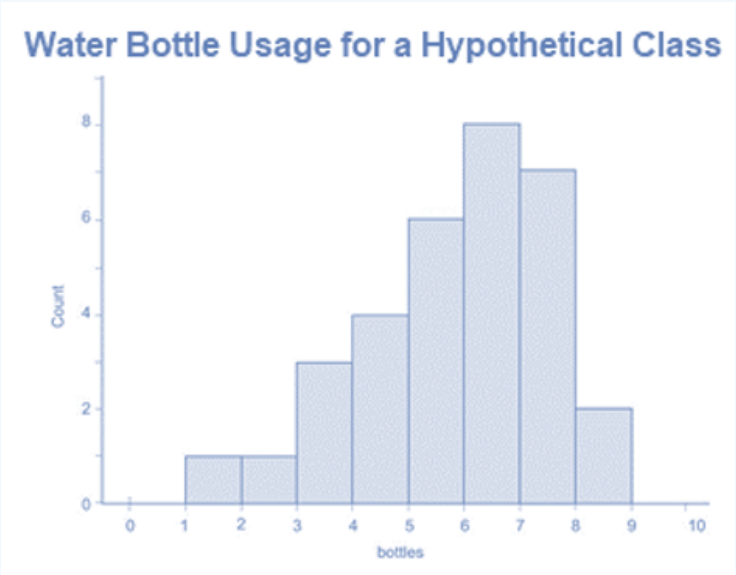
The spread of a distribution describes how far the data extend from the center. Common measures include the range, interquartile range (IQR), and standard deviation. Shape affects how we interpret these measures.

- In symmetric distributions, the spread is well-summarized by the standard deviation because deviations from the mean are balanced on both sides.
- In skewed distributions, the standard deviation may be inflated by extreme values in the tail. In these cases, the IQR provides a more stable description of spread because it focuses on the middle 50% of the data.

Understanding spread helps us determine whether the data are tightly clustered or widely dispersed, which in turn affects how confidently we can describe typical behavior.

Example 3.2.7

Refer to the histogram for the water bottle usage data.



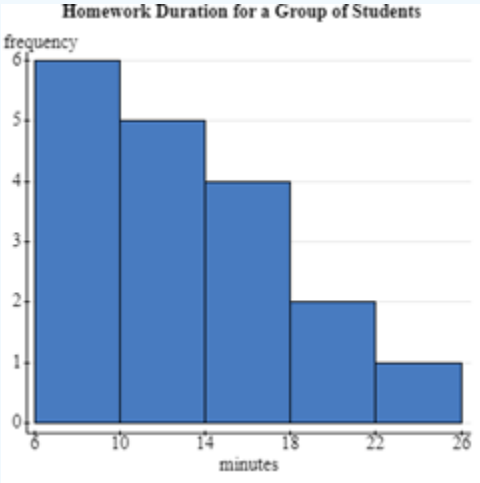
Describe the shape of distribution. How will the mean and median compare numerically? Which measures of center and spread will best represent the distribution?

Solution

The concentration of values between 5 and 8 bottles forms a clear unimodal peak. The tallest bar represents those students who used 6 bottles so 6 is the mode. The left tail extends farther than the right one, indicating left skew. From this shape alone, we can infer that the median will lie near the center of the main cluster while the mean will be pulled left by the longer tail. Therefore, the mean is smaller than the median. Due to the skewness, the median along with the IQR best describe the data distribution.

Example 3.2.8

Refer to the histogram for Homework Duration data.



Describe the shape of distribution. How will the mean and median compare numerically? Which measures of center and spread will be represent the distribution?

Solution

The concentration of homework times near on the left side of the histogram forms a clear unimodal peak. The tallest bar represents those students who spent from 6 up to 10 minutes. When data have been grouped into intervals, we cannot identify a specific mode. However, we can say the modal class is [6,10). The right tail stretches noticeably farther than the left one, indicating a skewed-right shape. From the graph alone, we can infer that the median lies near the center of the main cluster, while the mean is pulled to the right by the longer tail. As a result, the mean is larger than the median. Because of this skewness, the median together with the IQR provides the most reliable description of the overall distribution.

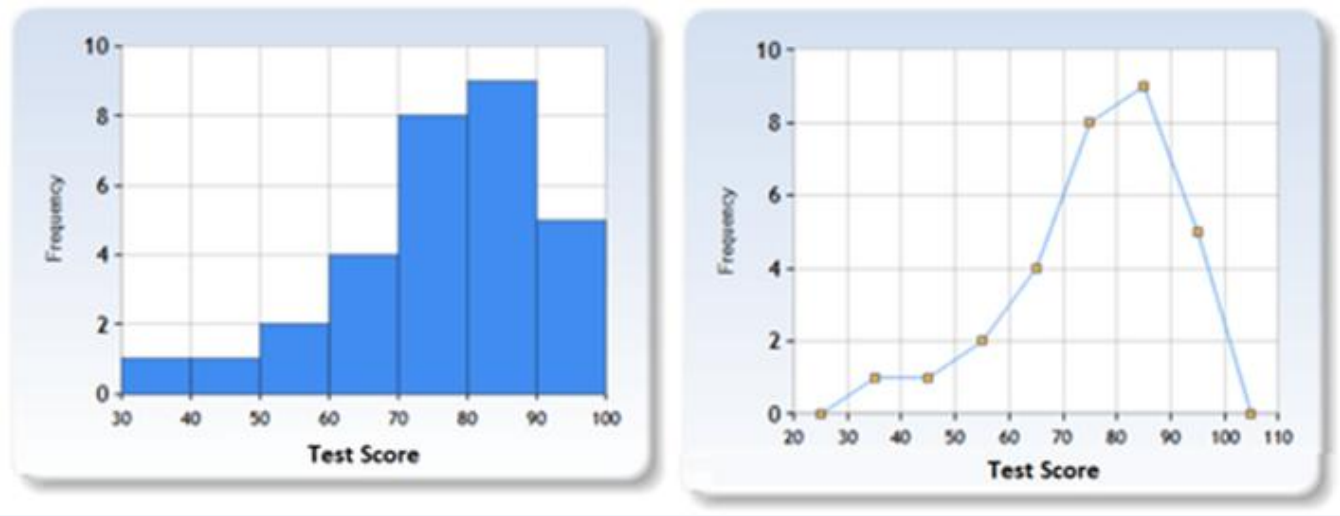
A grouped histogram is a powerful visual tool, but it does not preserve the exact raw data values. Because the data are aggregated into class intervals, we lose the precise information needed to compute exact mean, median, mode, range, IQR, standard deviation, and other descriptive statistics. What we see instead are counts of how many observations fall within each interval. This grouping is helpful for identifying overall patterns, but it prevents us from knowing the exact value of each individual data point.

Even so, grouped data still allow us to produce useful estimates. The mode can be estimated by giving the **modal class** - the class interval with the highest frequency. The median can be estimated by locating the class interval where the halfway point of the cumulative frequency occurs. The range can be approximated by using the midpoints of the first and last intervals, and the IQR can be estimated by identifying the intervals containing the first and third quartiles. These estimates are not exact, but they are often close enough to reveal the distribution's general behavior.

This is why histograms are so valuable: even without exact raw data, they allow us to infer the shape of the distribution, identify its approximate center, and gauge its spread. For many practical purposes—especially in introductory statistics -- these estimates provide meaningful insight into how the data are distributed and how best to summarize them.

Example 3.2.9

Ms. Winter made a histogram and frequency polygon of the science test scores from 5th period. Analyze the performance of this class by describing the shape, center, and spread of the distribution of test scores.



Solution

The histogram and frequency polygon display a single prominent peak, indicating a unimodal distribution. In fact, the modal class is [80, 90) points indicating that more students earned B's than any other grade. This peak appears toward the right side of the graph, while the tail stretches to the left. Because the longer tail is on the left, the distribution is skewed left rather than symmetric. This shape suggests that the majority of students scored in the higher score ranges, with fewer students earning lower scores.

To determine the median, note that there are $1+1+2+4+8+9+5=30$ test scores in all, so the median lies between the 15th and 16th ordered observations. These positions fall within the class interval [70, 80). Because the distribution is skewed left, the mean score is less than the median. The median provides a more stable measure of center for this distribution due to the skewness.

Although the exact minimum and maximum scores are not visible due to the grouping of the data, we can estimate the spread using the midpoints of the first and last class intervals. The approximate minimum is near 35, and the approximate maximum is near 95, giving an estimated range of about $95-35=60$ points. For a skewed distribution, the IQR is typically the most reliable measure of spread, as it focuses on the middle half of the data and is resistant to the influence of the long tail.

Section Summary

A frequency table organizes quantitative data into classes and records how many observations fall within each class; this count is the frequency. Classes may represent single values or equal-width intervals called bins. The relative frequency expresses these counts as percentages of the total.

A histogram is a graphical display of a frequency or relative frequency table, showing how data are distributed across the bins. A frequency polygon plots the midpoint of each bin at its frequency and connects the points with line segments, making it useful for visualizing the underlying density curve and for comparing multiple data sets. An ogive, constructed from cumulative frequencies, shows how data accumulate across the range of the variable.

Every distribution should be described in terms of shape, center, and spread. A single-peaked distribution is unimodal and may be symmetric or skewed. In a left-skewed distribution, most data lie on the right with a tail extending left. In a right-skewed

distribution, most data lie on the left with a tail extending right. Symmetry and skewness determine how the mean and median compare. In symmetric distributions they are approximately equal, while in skewed distributions the mean is pulled toward the longer tail (smaller than the median for left skew, larger for right skew).

Although exact descriptive statistics cannot be obtained from a histogram alone, important measures such as the mode, median, and range can be reasonably estimated from the graph, allowing meaningful interpretation even without seeing the raw data.

Section 3.2 Review Questions

1. Lois gathered her own data on the plastic water bottle use of her classmates, but she ran out of time as class was ending. When she arrived home, something had spilled in her backpack and smudged the data for the 2's. Fortunately, none of the other values were affected, and she knew there were 30 students in the class. Complete her frequency table.

Frequency Table for Lois's Water Bottle Data

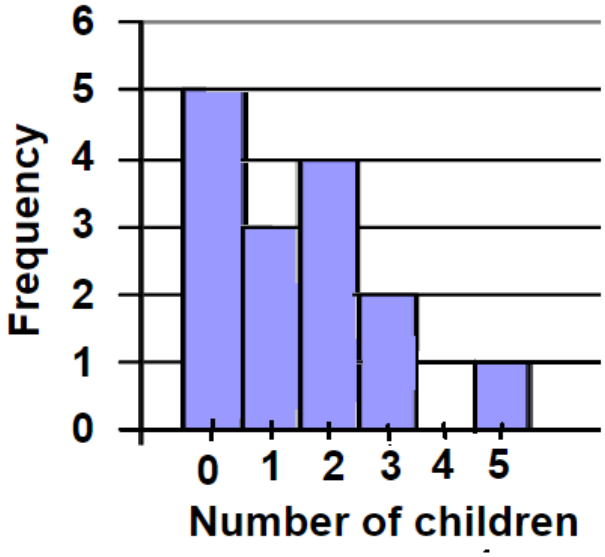
Number of Water Bottles Used per Week	Tally	Frequency	Relative Frequency
1	| |		
2			
3	| | |		
4	| |		
5	| | |		
6	𝍸 | |		
7	𝍸 |		
8	|		

2. A group of adults where asked how many cars they had in their household. The data are shown.



   a. Construct an ungrouped frequency table for data.
   b. Based on the frequency table, draw a histogram of the data.

3. A group of adults were asked how many children they have in their families. The histogram of ungrouped data shows the responses.



a. How many adults where questioned?
b. What percent of the adults questioned had 0 children?
c. What is the mode of this data set?

4. A wildlife researcher recorded the number of birds visiting a backyard feeder each morning for several weeks. She organized the counts into the grouped frequency table below. Use the table to answer the questions that follow.



a. What is the total number of data values in this distribution?
b. What class width (bin size) is used to group the data?
c. What is the modal class?
d. If an additional class were added to the end of the table, what would its lower and upper class limits be?

5. The following data represent ages of passengers on a train. Create a frequency table and a histogram of the data beginning the first class interval at 0 with bin width of 15.

35	2	45	40	48	4	3	22	40	27	37
42	24	60	44	84	2	20	26	51	48	58
38	27	38	1	38	48	6	17	62	35	21
57	36	40	44	20	58	40	18	31	27	

6. The SAT test scores for Michigan students in reading for several years are given in the table below (College Board: Michigan, 2012).

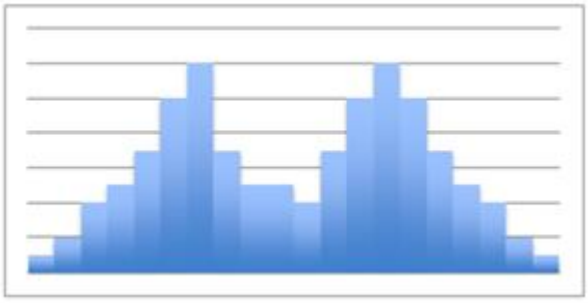
SAT Scores for Michigan Students

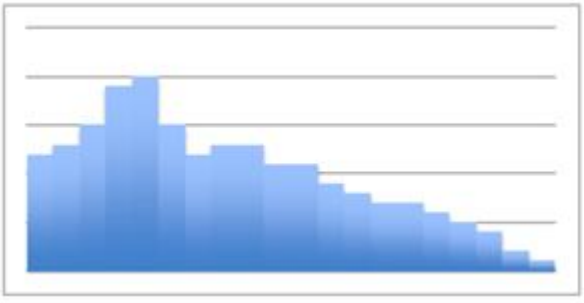


a. Create a grouped frequency table using 495-499 as the first class. Show frequencies, relative frequencies, and class marks.
b. Based on the frequency table, create a histogram of the data.

7. Describe the shape of each distribution shown below. Use the words *uniform, unimodal*, *bimodal*, *symmetric*, *skewed left*, or *skewed right*.

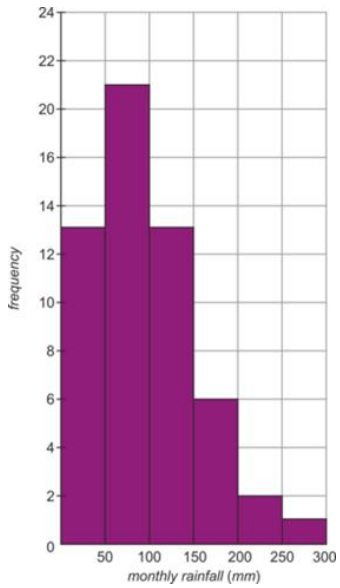
a.


b.


c.


d.


8. Monthly rainfall (in millimeters) for Beaver Creek, Oregon, was collected over a five-year period. Use the histogram of the data to answer the questions.



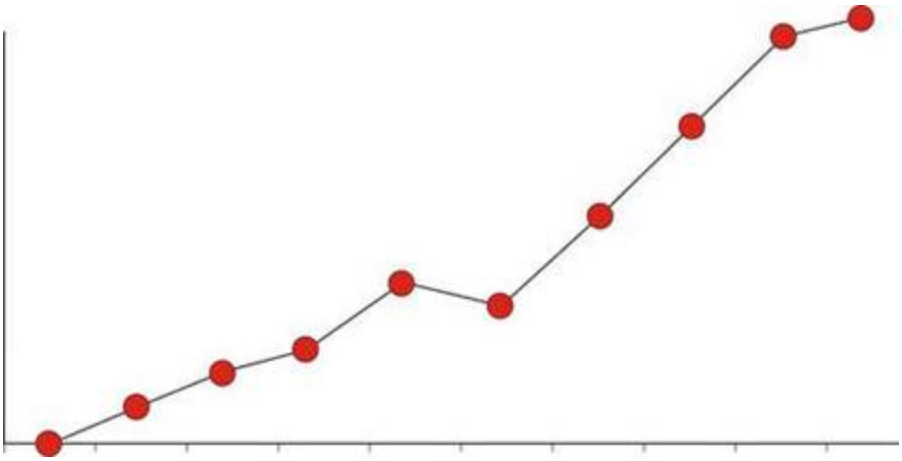
   a. Describe the shape of the data distribution.
   b. What bin size (class size) was used to create the histogram?
   c. In how many of these months was the total rainfall 150 mm or more?
   d. Suppose we make a relative frequency histogram. What is the relative frequency for the first class interval [0, 50) mm?
   e. In which interval of monthly rainfall will the median fall?
   f. Based on the shape of the histogram, which will be larger – the mean or median?
   g. Give an estimate for the range of monthly rainfall in Beaver Creek.

9. What do you think is the shape of the density curve of rolling a 6-sided die 1,000 times is? Why?

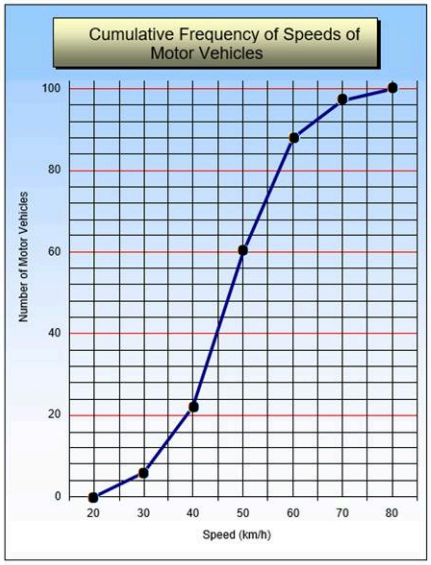
   - symmetric and uniform
   - skewed left
   - skewed right
   - symmetric and unimodal
   - symmetric and bimodal

10. Suppose a set of data is unimodal. Based solely on the mean and median given, determine the most probable shape of the distribution:
   a. Mean = 100; Median = 85
   b. Mean = 20; Median = 41
   c. Mean = 850; Median = 850

11. Tianna made the graph below for a homework question in her statistics class, but she forgot to label the graph or the axes and couldn't remember if it was a frequency polygon or an ogive plot. Which is it? How do you know?



12. The speeds several vehicles passing a certain point in a busy street are recorded and graphed as a histogram. Then, an ogive was created using the histogram. The ogive below shows the speed in km/h and the number of vehicles. Use the ogive to answer the questions.



a. How many vehicles are shown in this ogive?
b. How many vehicles were driving slower than 40 km/hr?
c. How many vehicles were driving between 50 km/hr up to 60 km/hr?
d. About 60 of the vehicles were driving slower than _____ km/hr.
e. In which class interval (bin) of speeds were there the smallest number of vehicles?
f. In which class interval (bin) of speeds were there the largest number of vehicles?

This page titled 3.2: Histograms, Frequency Polygons, and Ogives is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by Andy Jones.

- 11.7: Exercises by David Lippman & Jeff Eldridge is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

3.3: Other Quantitative Graphs and Data Plots

Learning Objectives

- Construct graphs of quantitative data including a dot plot, stem-and-leaf plot, and a line graph.
- Interpret information from graphs of quantitative data including a dot plot, stem-and-leaf plot, and a line graph.
- Describe distributions of univariate data in terms of shape, center, spread, and outliers.

Introduction

As we move beyond categorical graphs and histograms, we turn to several displays designed specifically for ungrouped quantitative data, where each individual value remains visible. Dot plots, stem-and-leaf plots, and line graphs preserves the raw data while still revealing patterns such as clusters, gaps, and unusual observations. These graphs are especially useful when the data set is small to moderate in size, making it possible to see both the overall shape of the distribution and the exact values that form it. In this section, we introduce these tools and learn how they help us display, compare, and interpret quantitative data with clarity and precision.

Dot Plots

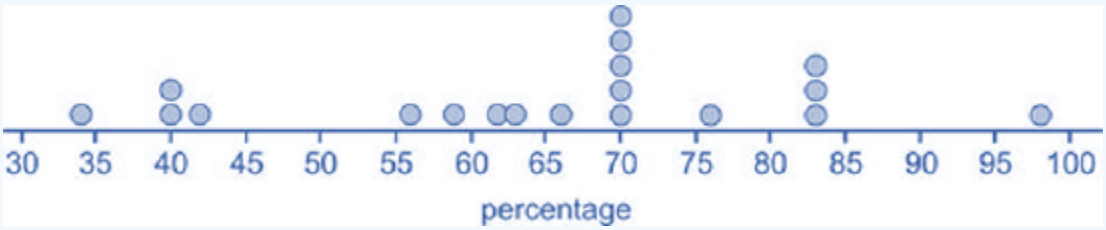
A **dot plot** is one of the simplest and most transparent ways to display a single quantitative variable. After selecting an appropriate numerical scale along a horizontal axis, each data value is plotted as a dot (or occasionally as an X) directly above its position on the number line. When a value occurs more than once, the dots are stacked vertically, creating a quick visual sense of frequency. Because every observation remains visible, dot plots preserve the raw data while still revealing important features such as clusters, gaps, and unusual values. Equal spacing along the axis is essential so that the shape, center, and spread of the distribution can be interpreted accurately.

Example 3.3.1

A soil-testing lab measured the moisture content (in percent) of samples taken from 19 locations across a research field.

$$\begin{array}{cccccc} 34 & 70 & 83 & 40 & 56 & 70 \\ 83 & 62 & 98 & 40 & 70 & 63 \\ 76 & 42 & 59 & 83 & 66 & 70 \\ 70 & & & & & \end{array}$$

Here is a dot plot of this information:



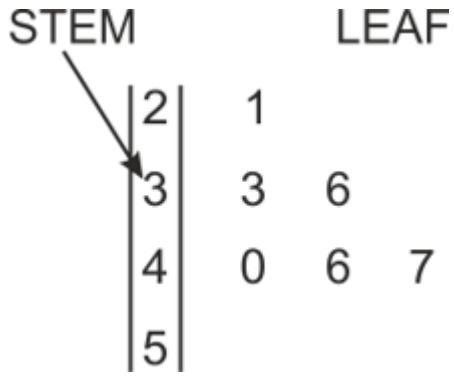
There are $n = 19$ data values, and the dot plot clearly shows the distribution centered at the 10th ordered value, giving a median of 70%. This is also the mode. The values are spread from 34% to 98%, producing a range of roughly 64 percentage points. The overall shape appears slightly left- skewed suggested by the longer stretch of values below the median.

Compared with a grouped histogram, dot plots have the advantage of showing each individual data point rather than grouping values into bins. This makes them ideal for identifying exact values, repeated observations, and fine-grained patterns in the distribution. However, dot plots become less practical when the data set is large or when the range of values is wide, as the display can become crowded or difficult to read. In such cases, grouped displays like histograms, frequency polygons, or ogives may be more effective.

Stem-and-Leaf Plots

A **stem-and-leaf plot** is another compact way to display quantitative data while preserving every original value. Unlike a grouped histogram, which groups observations into bins and hides the exact numbers, a stem-and-leaf plot groups the data into bins while keeping the raw data visible and sorted. This makes it especially useful for identifying the shape, center, spread, and outliers of a distribution while still retaining full numerical detail.

In a stem-and-leaf plot, each data value is separated into two parts: a *stem*, which contains the leading digit(s), and a *leaf*, which contains the remaining digit. For example, if the tens place is used as the stem, then the values 21, 33, 36, 40, 46, and 47 would appear as:



This display sorts the data into intervals of width 10 while showing every individual value. Because the leaves appear in order, it becomes easy to locate the median, identify the mode, and determine the five-number summary. Stem-and-leaf plots also make the distribution's shape visible—clusters, gaps, and skewness stand out clearly.

Example 3.3.2

While traveling on a long train journey, Rowena recorded the ages of all the passengers in her train car. The ages she collected are shown below.

35	42	38	57	2	24	27	36	45	60
38	40	44	1	44	48	84	38	20	4
2	48	3	20	6	40	22	26	17	18
40	51	62	27	48	35	27	37	58	21
40	58	31							

Construct a stem-and-leaf plot for the data, then use the plot to determine the median and mode ages.

Solution

The first step is choosing an appropriate stem. Since the ages range from 1 to 84, the tens digit makes a sensible stem, running from 0 through 8. This allows values such as 2 to be written as a leaf of 2 on the 0 stem, and 84 as a leaf of 4 on the 8 stem. Filling in the stems with the corresponding leaves produces the initial stem-and-leaf plot.

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

From this plot, it is already clear that the 40–49 interval contains the most passengers and is the modal class. To accurately determine the median and mode, it is helpful to rewrite the plot with the leaves on each stem arranged in ascending order. When ordering the leaves, note that some stems may have no values. For example, if no ages fall in the 70–79 range, the stem "7" should appear with no leaves. Empty stems must still be included to preserve spacing and maintain an accurate picture of the distribution.

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

Once the ordered plot is complete, the mode becomes obvious: the stem "4" contains four leaves of 0, corresponding to the repeated age 40. Thus, the mode is 40. To find the median, count the total number of data values. There are $n = 43$ ages, so the median is the 22nd value in the ordered list. Reading across the ordered stem-and-leaf plot, the 22nd value is 37. Therefore, the median age is 37.

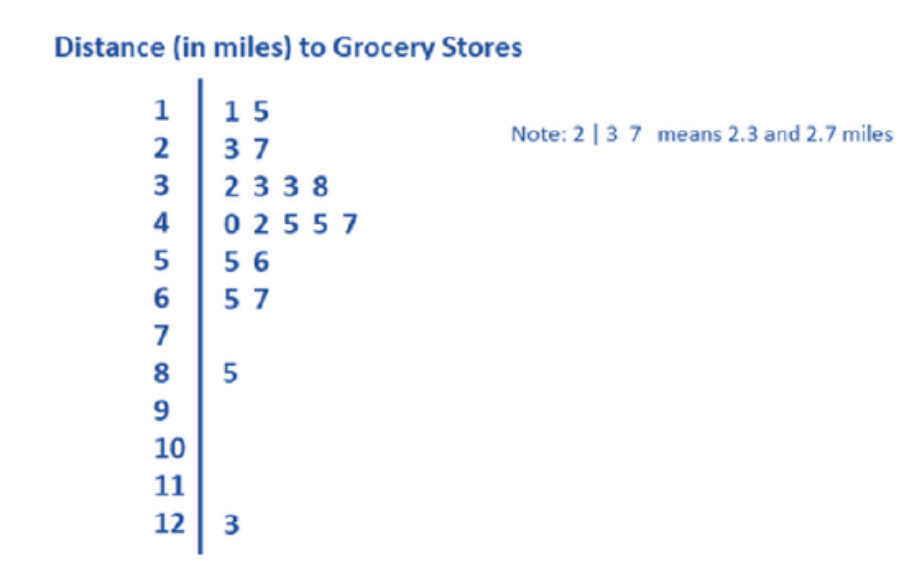
It is important to clarify to a reader how to interpret the data in a stem-and-leaf plot, especially for someone viewing the display without access to the original data. A brief note beneath the graph in the previous example -- such as “8 | 4 means 84 years old” -- ensures that the stems and leaves are interpreted correctly. If the data include one-decimal-place measurements such as 1.1, 2.7, or 6.5, the stem might represent the ones place and the leaf the tenths place. Here is an example:

Example 3.3.3

Data were collected on the distances (in miles) from home to several nearby supermarkets. They are already ordered for convenience.

$$\begin{array}{cccccc} 1.1 & 1.5 & 2.3 & 2.7 & 3.2 & 3.3 \\ 3.8 & 4.0 & 4.2 & 4.5 & 4.5 & 4.7 \\ 5.5 & 5.6 & 6.5 & 6.7 & 8.5 & 12.3 \end{array}$$

To construct a stem-and-leaf plot for these values, the stem represents the ones place (ranging from 1 through 12) and the leaf represents the tenths place. Because the stems now correspond to whole miles and the leaves to decimal parts, it is essential to include a note such as “2 | 3 means 2.3” so the values are interpreted correctly and not mistaken for whole numbers like 23.



Stem-and-leaf plots are most effective when the data set is moderate in size and the values fall within a manageable range. When observations contain many digits, include decimals, or span widely separated intervals, the stems may need to be adjusted or truncated to keep the display readable. Conversely, when the data are tightly clustered -- such as a set of single-digit counts -- nearly all leaves may fall on the same stem, offering little insight beyond what a simple list would show. In these situations, other displays such as dot plots or histograms often provide clearer structure. Still, when the data are well-suited to the format, stem-and-leaf plots offer a rare combination of numerical precision and visual clarity.

Line Graphs

A **line graph** -- also called a **time-series plot** -- is used when the order in which data are collected matters. Unlike histograms, dot plots, or stem-and-leaf plots which display values sorted from smallest to largest, a line graph preserves the sequence of

observations. This makes it ideal for showing how a variable changes over time and for identifying trends, cycles, or abrupt shifts in behavior.

In a line graph, the horizontal axis represents time (years, months, days, or any other chronological unit), and the vertical axis represents the measured variable. Each observation is plotted as a point at its corresponding time value. Successive points are then connected with straight line segments. These segments reveal the rate of change between observations—steeper segments indicate rapid increases or decreases, while flatter segments show periods of stability.

A line graph can be viewed as a connected scatterplot, where the connections add context about how the data evolve. This makes it easier to spot long-term trends, seasonal patterns, or unusual deviations that might warrant further investigation.

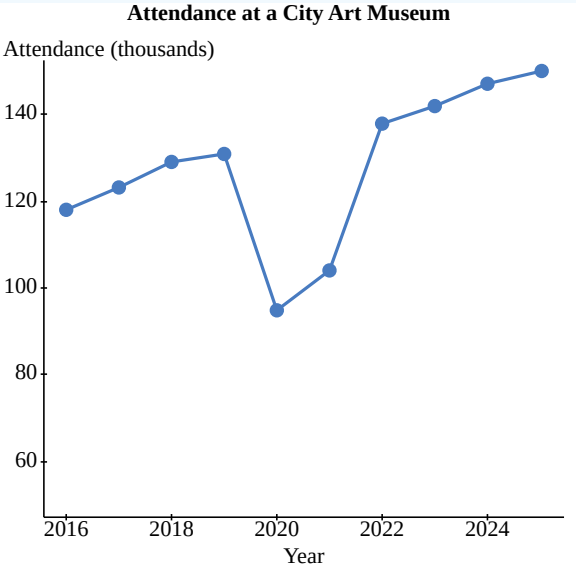
Example 3.3.4

Suppose a city's art museum recorded its annual visitor attendance (in thousands) over a decade. Analyze the yearly attendance for patterns and trends using by making a line graph.



Solution

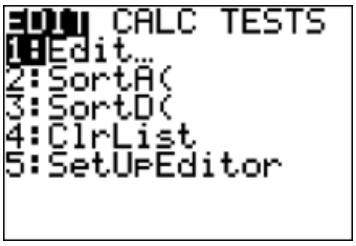
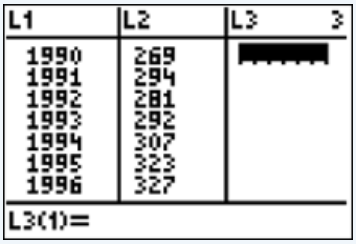
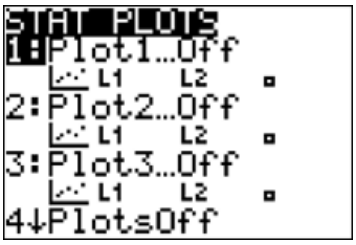
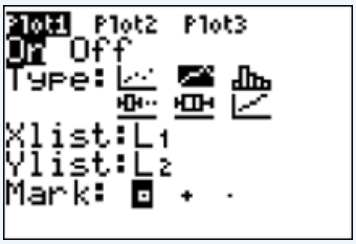
To create a line graph, place the years 2010–2019 on the horizontal axis and the attendance values on the vertical axis. Plot each year's attendance as a point, then connect successive points with line segments to show how attendance changed from year to year.

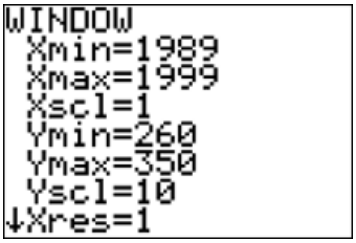
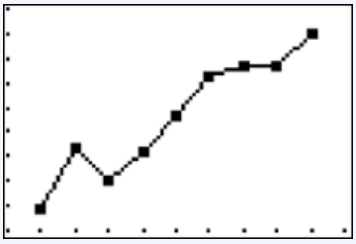


Once graphed, several patterns become clear:

- Attendance drops sharply in 2020, shown by the steep segment sloping downward.
- The years following 2020 show a steady recovery, with strong increases through 2025.
- The largest single-year increase occurs between 2021 and 2022.
- The overall trend from 2016 to 2025 is upward, despite the temporary decline.

Making a Line Graph on the TI-83/84 Calculator

1.	Press [**STAT**] and then choose **1:Edit** to input the data into two lists.	
2.	Here, the year is input into L1 and the waste is input into L2. Make sure the year and its corresponding data value are matched.	
3.	Press **[2ND] [Y=]** to enter the **STAT PLOTS** menu. Choose **Plot 1**.	
4.	Move the cursor to the word **On** and press **[ENTER]** to turn the plot on. Change the **Type** of plot to a line graph (the second icon), and choose the time list (L1) for **Xlist.** Choose the data list (L2) for **Ylist**. Decide which mark you want to show for each data value.	

5.	We need to set a window. Press **[WINDOW]** and enter an appropriate window to display the plot. Let's start the year one year earlier and go one year further (1989 to 1999.) The values appearing on the vertical axis range from 269 to 340, so will will begin at 260 and end at 350, scaling by 10.	
6.	Press **[GRAPH]** to display the time series graph. If you press **[TRACE]** and then use the left or right arrows to see the actual data values at each year. **Note:** Instead of setting the WINDOW manually, you can let the calculator set its own window by pressing **[ZOOM]** and then choosing **9:ZoomStat**.	

Section Summary

Three common graphs used to represent quantitative data are the dot plot, the stem-and-leaf plot, and the line graph. A dot plot displays each data value as a dot along a number line, providing a quick visual impression of the distribution's shape, center, and spread. Dot plots are easy to interpret but can become cumbersome for very large data sets.

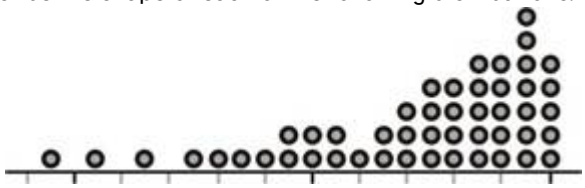
Stem-and-leaf plots show similar information while preserving every original data value. Because the stems group values and the leaves display the exact digits, these plots make it easy to identify the mode, median, and overall distribution.

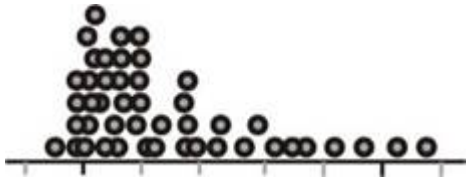
A line graph is used when the chronological order of the data matters. By plotting values over time and connecting them with line segments, a line graph highlights increases, decreases, and cyclical patterns. This makes it especially useful for identifying trends and making informed predictions about future behavior.

Together, these graphs provide complementary ways to visualize quantitative data, each suited to different types of questions and patterns in the data.

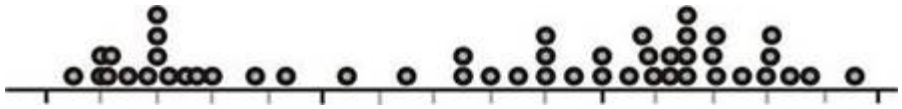
Section 3.3 Review Questions

1. Describe the shape of each of the following distributions.

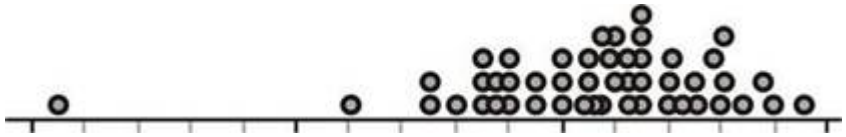
a. 

b. 

c.



d.



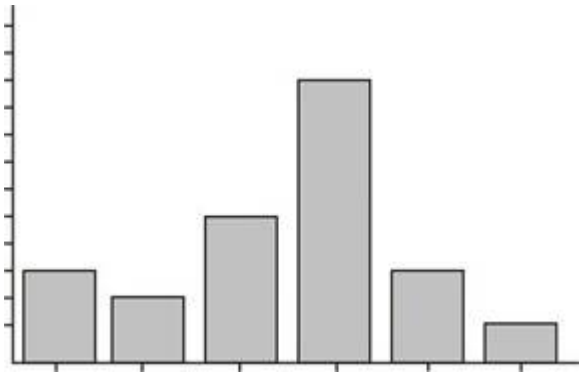
2. The dot plot shows the distribution of the number of books each student in Mr. Herman's class read last summer. Use it to answer the questions.



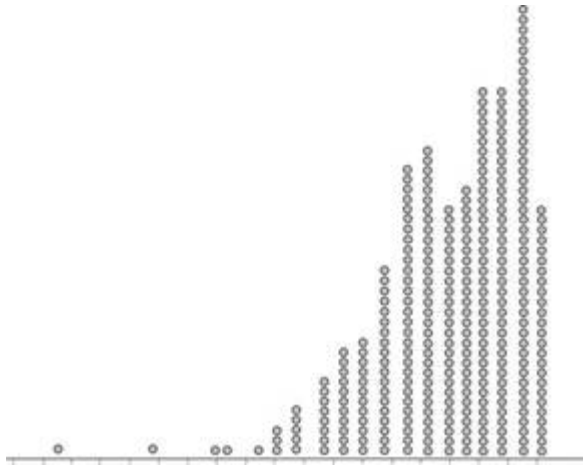
   a. Describe the shape of the distribution.
   b. Find the mean.
   c. Find the median.
   d. Find the mode.
   e. Find the range.
   f. Find the IQR.
   g. Determine if any students are outliers in terms of the number of books they read.

3. Match the distribution of the real-world data distribution with its probable graph.
   a. Andy collected and graphed the heights of all the adults at his church.
   b. Brittany asked each of the students in her statistics class to bring in 20 pennies selected at random from their pocket or piggy bank. She created a plot of the dates of the pennies.
   c. Thamar asked her friends what their favorite movie was this year and graphed the results.
   d. Juno bought a large box of doughnut holes at the local pastry shop, weighed each of them, and then plotted their weights to the nearest tenth of a gram.

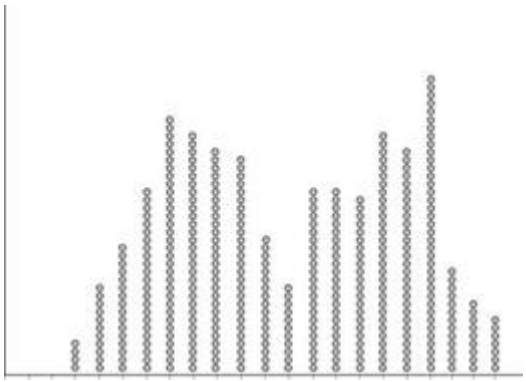
I.



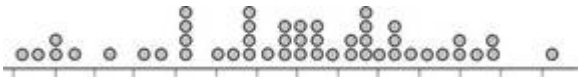
II.



III.

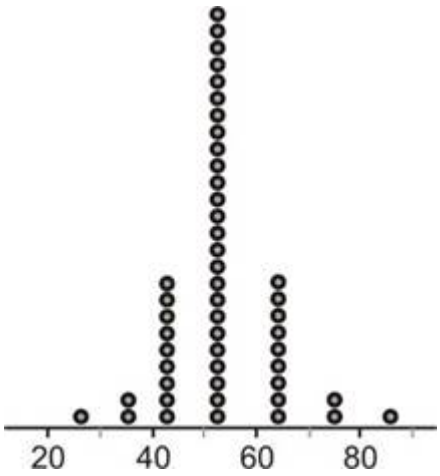


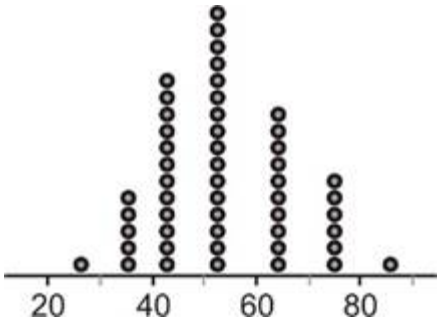
IV.

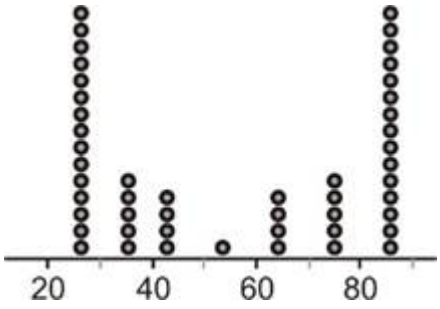


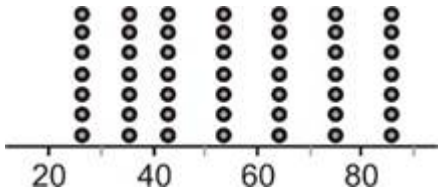
4. Use the graphs below to answer each part.
   a. Identify the overall shape of each distribution.
   b. Which measure of center (mean or median) would be best to characterize the centers of these distributions?
   c. Which of these distributions has the smallest standard deviation?
   d. Which of these distributions has the largest standard deviation?

I.

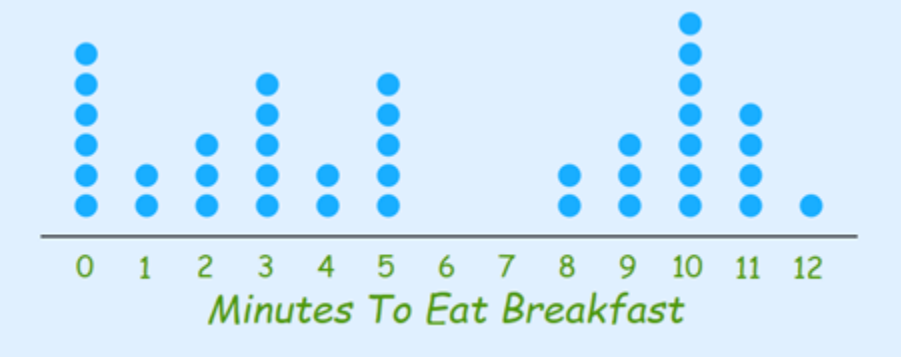


II. 

III. 

IV. 

5. The dot plot below shows the results of a survey when several people were asked "*How long does it take you to eat breakfast*?" Answer each question. Then, identify the level of graph comprehension that it addresses as *reading the data, reading between the data*, or *reading beyond the data*.



   a. How many people said they eat breakfast in 3 minutes?
   b. What is the most amount of time any person uses to eat breakfast?
   c. How much time do people use to eat breakfast most often?
   d. What fraction of the people eat breakfast in less than 5 minutes?
   e. How many people use 480 seconds to eat breakfast?

6. Here is a data set summarized as a stem-and-leaf plot. Assume the stem represents the tens place and the leaf represents the ones place.

   6 | 0023667888

   7 | 0001133556677899

8 | 05679

9 | 068

a. How many data values are in this data set?
b. What is the minimum value in the last class?
c. What is the minimum value in the entire set of data?
d. How many of the original values are greater than 70?
e. What is the relative frequency of values less than 65?
f. What is the range of the entire set of data?
g. How can the shape of the distribution of these data be described?
h. Without calculation, which should be larger – the mean or median

7. Use the following stem-and-leaf plot which shows data collected for the speed of 40 cars in a 35-mph limit zone in Culver City, California.

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

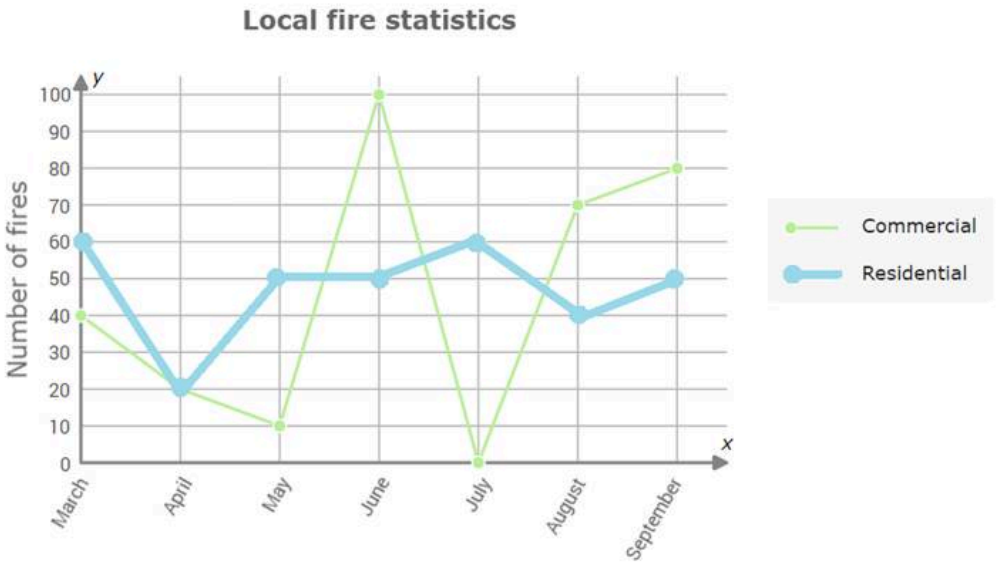
a. Find the mean, median, and mode speed.
b. Create a frequency table, starting at 25 mph with a bin width of 5 mph.
c. Use the table to draw a histogram with the intervals from the frequency table in part (b.)

8. Just like presidents of the United States, Canadian prime ministers must be sworn into office. The following data represents the ages of 22 Canadian prime ministers when they were first sworn into office. Construct a stem-and-leaf plot to represent the ages. Once you have made the plot, find the median and mode from the plot.

52	74	60	39	65	46	55	66
54	51	70	47	69	47	57	46
48	66	61	59	46	45		

9. The following line graphs show the number of calls that a city fire department made from March through September. Use the graph to answer the questions. Use reasonable estimates as needed.

a. How many commercial fires did the fire department report for August?



b. How many more residential fires than commercial fires were there in May?
c. During which month(s) (if any) were there no commercial fires?
d. During which month(s) (if any) were there no residential fires?
e. During which month(s) were there the same number of residential fires as commercial fires?
f. During which month(s) were there more commercial fires than residential ones?
g. Between which consecutive months was there the greatest change in the number of commercial fires?
h. Which appears to have more variation – the number of residential fires or the number of commercial fires?
i. How many fires were there in all in July?
j. During which month was there the fewest number of fires?
k. How many commercial fires were there in all from March through September?
l. What does the horizontal line segment between May and June indicate about the number of residential fires?

This page titled 3.3: Other Quantitative Graphs and Data Plots is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

3.4: Box-and-Whisker Plots

Learning Objectives

- Construct and interpret a modified and unmodified box-and-whisker plot.
- Describe the shape, center, spread, and outliers of data shown in a box-and-whisker plot.
- Interpret box-and-whisker plot information in context.
- Use side-by-side box plots to compare distributions.

Introduction

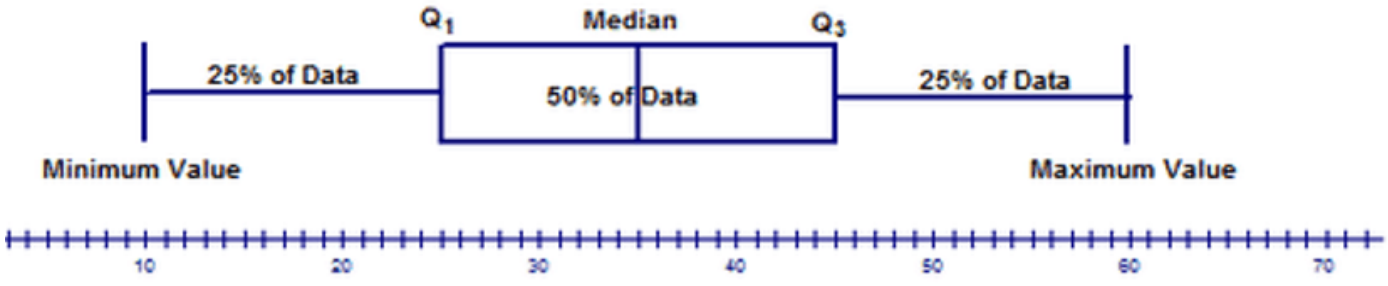
A box-and-whisker plot provides a compact visual summary of a quantitative data set by displaying its five-number summary -- the minimum, first quartile, median, third quartile, and maximum. These five values identify positions in the distribution that divide the data into four equal parts, each containing 25% of the observations. After working with categorical graphs, histograms, dot plots, line graphs, and stem-and-leaf plots, the box-and-whisker plot adds a new perspective by emphasizing how data are positioned rather than how they accumulate.

In this section, we will introduce the construction and interpretation of box-and-whisker plots and use them to study the fundamental features of a distribution: its shape, center, spread, and potential outliers. Because the plot highlights quartiles and extremes, it becomes a powerful tool for comparing multiple data sets side-by-side and for identifying unusual observations that fall far from the main body of the data.

Basics of a Box-and-Whisker Plot

A **box-and-whisker plot** (often shortened to *box plot*) is a graphical display that summarizes a quantitative data set using its five-number summary. These five values -- the minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum -- locate positions in the distribution that divide the data set into four parts, each containing 25% of the observations. By plotting these values on a number line and connecting them in a structured way, a box-and-whisker plot highlights how the data are centered, how they are spread out, and whether any portions of the distribution extend farther than others.

This display of the five-number summary produces a box-and-whisker plot as shown here:



A box-and-whisker plot is named for its two main components: a *box*, which spans from Q_1 to Q_3, and two *whiskers*, which extend

outward from the box to the minimum and maximum values. The box represents the middle 50% of the data and its width equals the interquartile range (IQR = $Q_3 - Q_1$). For example, if $Q_1 = 25$ and $Q_3 = 45$, then the middle half of the data stretches across an IQR of 20. A vertical line inside the box marks the median, dividing the box into two smaller regions, each containing 25% of the data.

The whiskers show the remaining halves of the distribution. The left whisker extends from Q_1 down to the minimum value, and the right whisker extends from Q_3 up to the maximum value. If the minimum is 10 and Q_1 is 25, then the lowest 25% of the data fall between 10 and 25. Likewise, if Q_3 is 45 and the maximum is 60, the highest 25% of the data fall between 45 and 60.

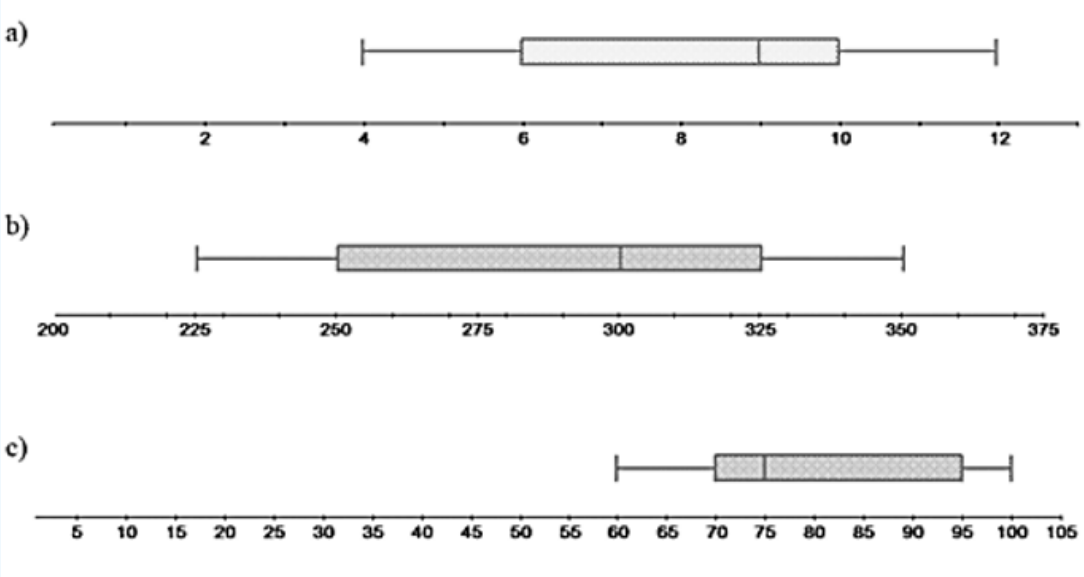
Features of the box and whiskers provide insight into the shape of a distribution:

- Median position: If the median lies near the center of the box, the distribution is roughly symmetric. A median left of center suggests a right-skewed distribution, while a median right of center suggests a left-skewed distribution.
- Whisker length: If the whiskers are about the same length, the distribution is roughly symmetric. A longer right whisker indicates right skew; a longer left whisker indicates left skew.

In the earlier example, the median appears centered within the box and the whiskers are similar in length, suggesting an approximately symmetric distribution. In such cases, the mean and median are typically close in value.

Example 3.4.1

For each box-and-whisker plot, list the five-number summary and describe the shape of the distribution based its center and spread.

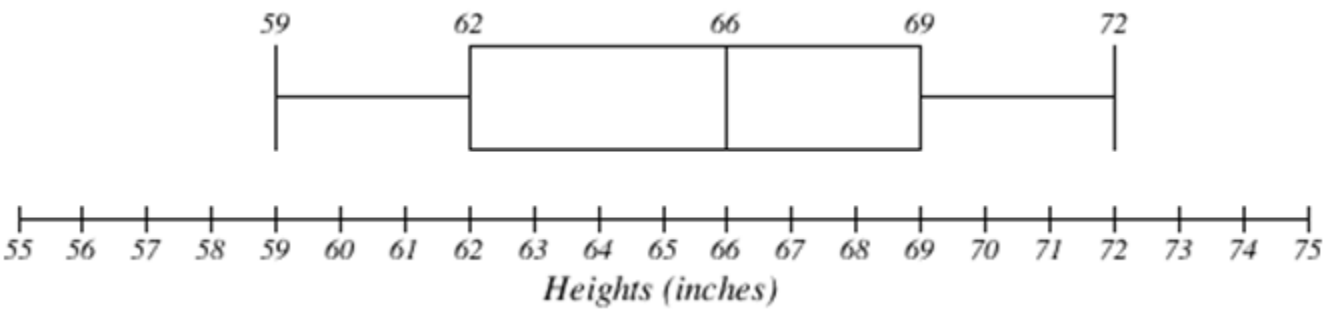


Solution

a. minimum = 4, Q1 = 6, median = 9, Q3 = 10, maximum = 12. The median lies to the right of the box's center, indicating a left-skewed distribution.
b. minimum = 225, Q1 = 250, median = 300, Q3 = 325, maximum = 350. The median lies to the right of the box's center, indicating a left-skewed distribution.
c. minimum = 60, Q1 = 70, median = 75, Q3 = 95, maximum = 100. The median lies to the left of the box's center, indicating a right-skewed distribution.

Example 3.4.2

The box plot represents the heights of a group of several females.



a. What is the median height for the females in this group?

Solution

The median is located at the vertical line inside the box of the box plot. Here, the median is 66 inches.

b. What is the interquartile range of the heights for the females in this group?

Solution

$IQR = Q_3 - Q_1 = 69 - 62 = 7$ inches.

c. What percent of the females in this group are 62 inches or shorter?

Solution

62 inches corresponds to Q_1. This value is also the 25th percentile, so 25% of the females are 62 inches or shorter.

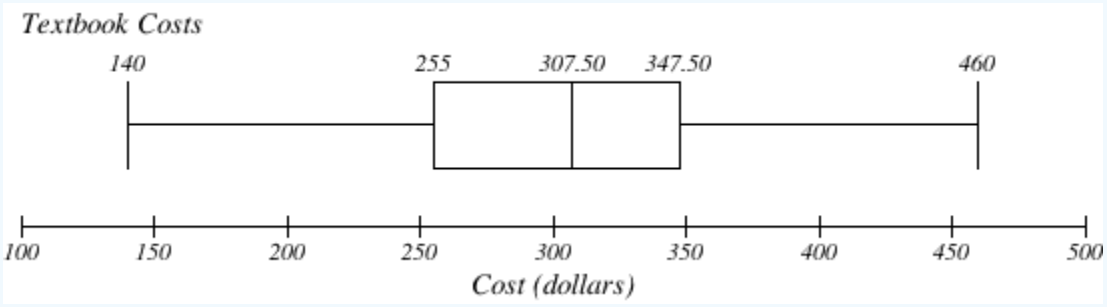
d. How tall are the tallest 25% of females in this group?

Solution

The tallest 25% of females would be taller than the value of Q_3 = 69 inches.

You Try It 3.4.A

The box plot shows the total cost of textbooks for the fall semester for a sample of PGCC students.



a. What is the most any student spent on textbooks for the semester?
b. What percent of students spent between $255 and $347.50 on textbooks for the semester?

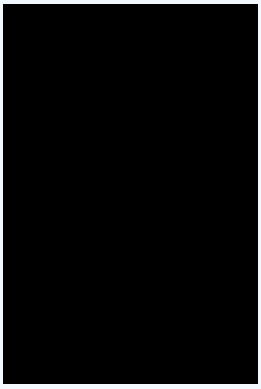
c. What percent of students spent \$347.50 or less on textbooks for the semester?
d. If there were $n = 80$ students in this sample, how many of them spent \$255 or less on textbooks?

Answer available at the end of this section.

Constructing a Box-and-Whisker Plot

Example 3.4.3

Suppose a school district is reviewing the percentage of students meeting a particular math benchmark across its ten elementary schools. The percentages for the most recent year are listed below:



Solution

To construct a box-and-whisker plot, begin by finding the five-number summary. Review Section 2.2 as needed.

First, order the data:

3 9 33 49 51 59 89 95 95 98

There $n = 10$ values, which is even. The median will be between the data values in positions #5 and #6:

3 9 33 49 **51 | 59** 89 95 95 98

$$\text{Median} = \frac{51+59}{2} = 55$$

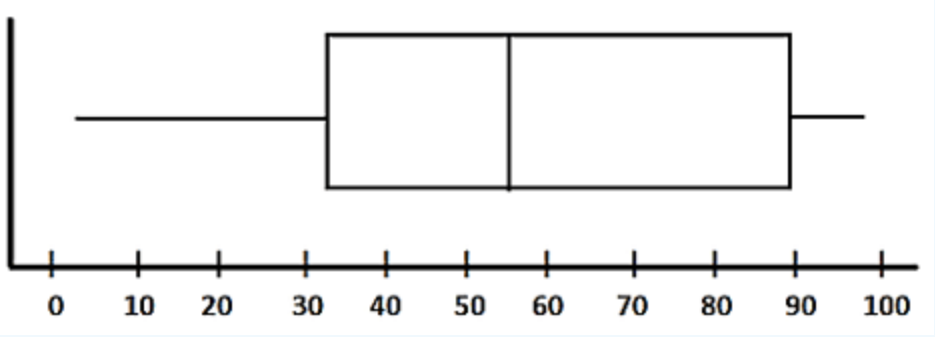
There are five values remaining in each half of the data set. Q_1 and Q_3 will each be at position #3 in each half:

3 9 **33** 49 **51 | 59** 89 **95** 95 98

$$Q_1 = 33 \; Q_3 = 95$$

The minimum is 3 and the maximum is 98, giving the five-number summary: [3, 33, 55, 95, 98].

To create the box portion of the plot, draw a rectangle extending from Q_1 (33) to Q3 (95) Place a vertical line inside the box at the median (55). Then attach whiskers from Q_1 (33) down to the minimum (3) and from Q3 (95) up to the maximum (98). This completes the box-and-whisker plot.



Modified Box-and-Whisker Plots

Occasionally, a box-and-whisker plot displays a whisker that is noticeably longer than the other. When this happens, we must decide whether the long whisker reflects genuine skewness in the data or whether it is caused by an outlier—a value that is unusually small or unusually large compared with the rest of the distribution. Because outliers occur at the extremes, they can stretch a whisker and make a distribution appear more skewed than it truly is. **Modified box-and-whisker plots** help us identify and display outliers clearly so that the plot is not misleading.

Example 3.4.4

An after-school program recorded how many minutes eleven children spent reading independently during one afternoon session. The reported reading times (in minutes) were

80 83 77 95 85 74 34 68 90 82 75

Solution

To construct a box-and-whisker plot, begin by finding the five-number summary.

First, order the data:

34 68 74 75 77 80 82 83 85 90 95

There $n = 11$ values, which is odd. The median will be between the data value in position #6:

34 68 74 75 77 **80** 82 83 85 90 95

$$\text{Median} = 80$$

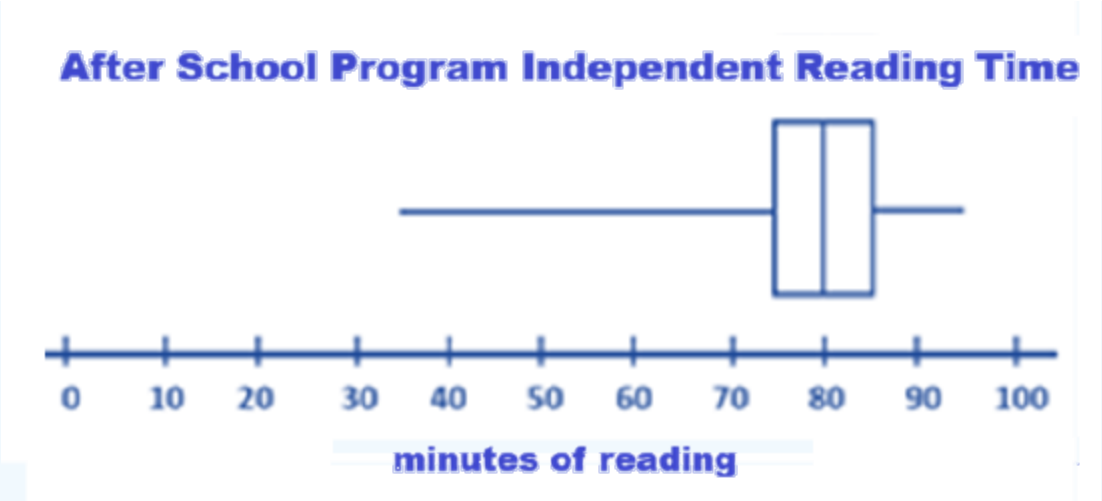
There are five values remaining in each half of the data set. Q_1 and Q_3 will each be at position #3 in each half:

34 68 **74** 75 77 **80** 82 83 **85** 90 95

$$Q_1 = 74 \; Q_3 = 85$$

The minimum is 3 and the maximum is 98, giving the five-number summary: [34, 74, 80, 85, 95].

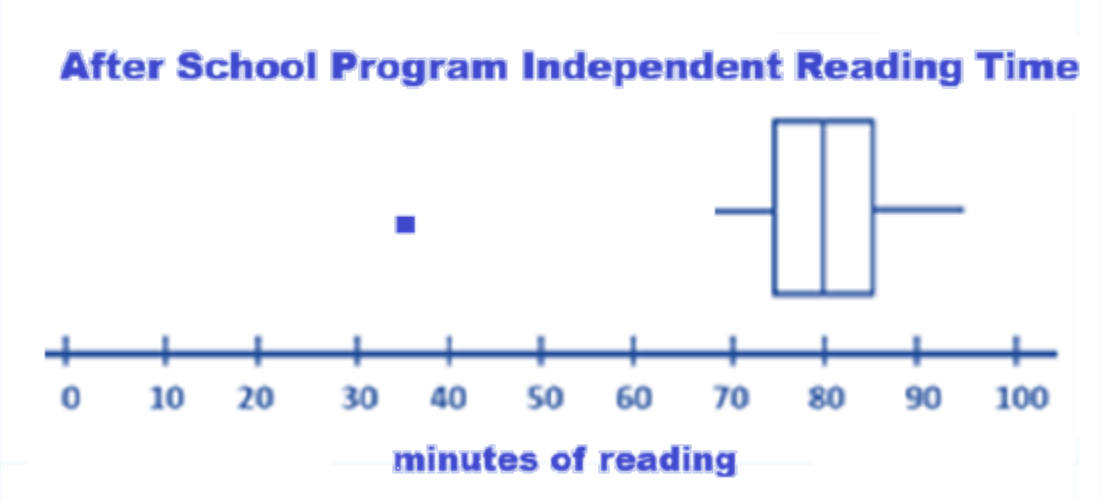
A standard box-and-whisker plot for these data is shown below.



The left whisker is noticeably longer than the right. To determine whether this is due to skewness or an outlier, apply the $1.5 \times IQR$ rule:

1. $Q_1 = 74$ and $Q_3 = 85$.
2. $\text{IQR} = Q_3 - Q_1 = 85 - 74 = 11$.
3. $\text{Step} = 1.5 \times \text{IQR} = 1.5 \times 11 = 16.5$.
4. $\text{Lower Fence} = Q_1 - \text{Step} = 74 - 16.5 = 57.5$ and $\text{Upper Fence} = Q_3 + \text{Step} = 85 + 16.5 = 101.5$.
5. Any values outside the fenced interval will be considered outliers. Consulting the data, 34 minutes is lower than the lower fence (57.5.) Therefore, it is an outlier. There are no data values higher than the higher fence (101.5.)

When outliers are present, a **modified box plot** is preferred. In a modified box plot, outliers are plotted as individual points. Whiskers extend only to the smallest and largest non-outlier values. In this example, the left whisker stops at 68 — the smallest value that is not an outlier — and the point at 34 is shown separately.



We can now better describe the distribution better. Instead of showing strong left-skewness, the distribution of reading times is roughly symmetric with a low outlier.

Caution: A common misconception is that whiskers should stop at the fence values. In fact, whiskers always stop at the nearest non-outlier data values.

Example 3.4.5

The scores for a women's golf team in tournament play are listed below. Construct a box-and-whisker plot to show the distribution. Use a modified box plot if necessary.

89 90 87 95 86 81 111 108 83 88 91 79

Solution

The data are ordered and the five-number summary is shown:

There are five values remaining in each half of the data set. Q_1 and Q_3 will each be at position #3 in each half:

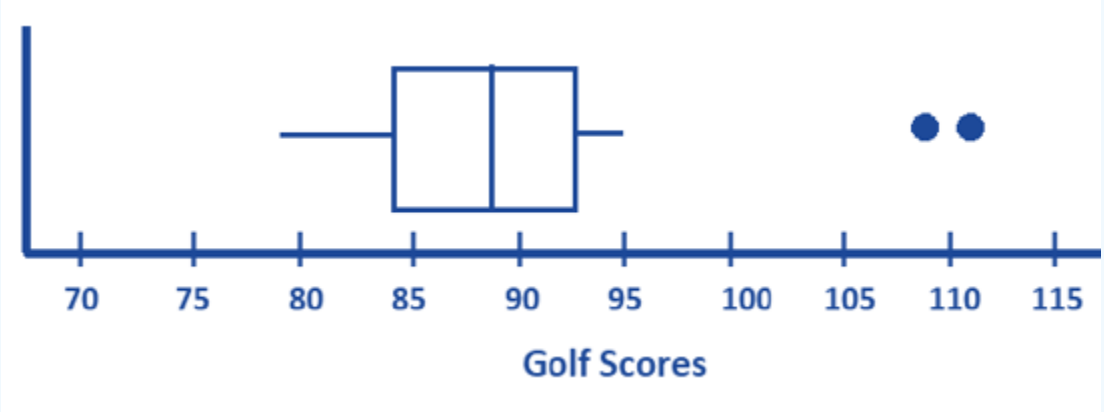
79 81 **83 | 86** 87 **88 | 89** 90 **91 | 95** 108 111

$$Q_1 = \frac{83+86}{2} = 84.5 \text{ Median} = \frac{88+89}{2} = 88.5 \; Q_3 = \frac{91+95}{2} = 93$$

The minimum is 79 and the maximum is 111, giving the five-number summary: [79, 84.5, 88.5, 93, 111].

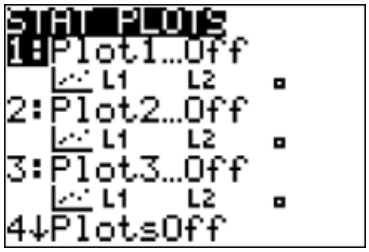
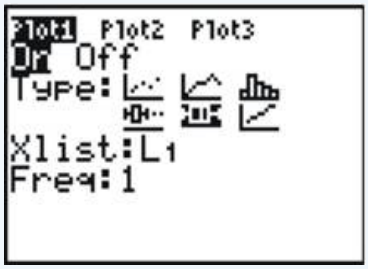
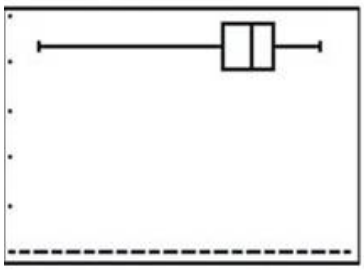
Before drawing the box-and-whiskers plot, check for outliers:

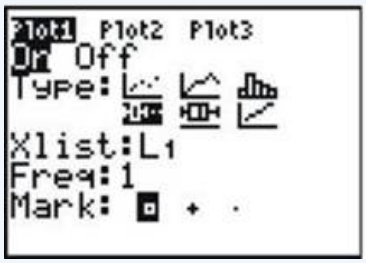
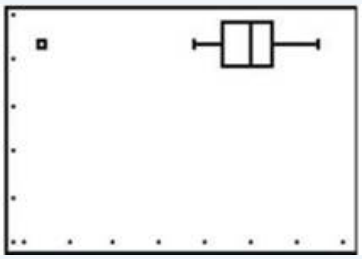
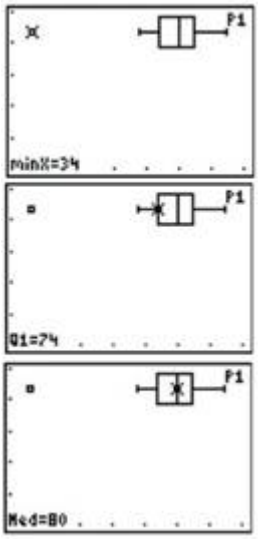
- $\text{IQR} = Q_3 - Q_1 = 93 - 84.5 = 8.5$.
- $\text{Step} = 1.5 \times \text{IQR} = 1.5 \times 8.5 = 12.75$.
- $\text{Lower Fence} = Q_1 - \text{Step} = 84 - 12.75 = 71.75$ and $\text{Upper Fence} = Q_3 + \text{Step} = 93 + 12.75 = 105.75$.
- Any data outside the interval [71.75,105.75] are outliers. In this case, there are two golfers who had outlier scores: 108 and 111



Based on the shape of the box-and-whiskers plot, the distribution of golf scores appears roughly symmetric with two high outlier scores.

There are a couple of cautions about making a box-and-whisker plot. One common misconception is that whiskers should stop at the fence values. In fact, whiskers always stop at the nearest non-outlier data values. Also, there may be no outliers, one outlier, or multiple outliers on either or both ends of the distribution. The only way to know is to use the outlier test. Because the procedure can be long and tedious with large data sets, we often can do a quick check for outliers and shape using the TI calculator:

1.	Enter the data into a list such as L1 as we have done before. Then choose **[2ND] [Y=]** to enter the **STAT PLOTS** menu.	
2.	Choose a plot (such as **Plot 1**) and change the settings as shown. Be sure to set the graph to **On**. Under **Type**, you will notice what looks like two different box and whisker plots. The second one is the regular (unmodified) box plot. Use your right arrow key to select it and press **[ENTER]**. Make sure the correct list of data is selected for **Xlist**.	
3.	Setting your own window is not as important for a box plot, so we will use the calculator's ability to automatically scale a window to our data. Press **[ZOOM]** and select **9:Zoom Stat.**	 

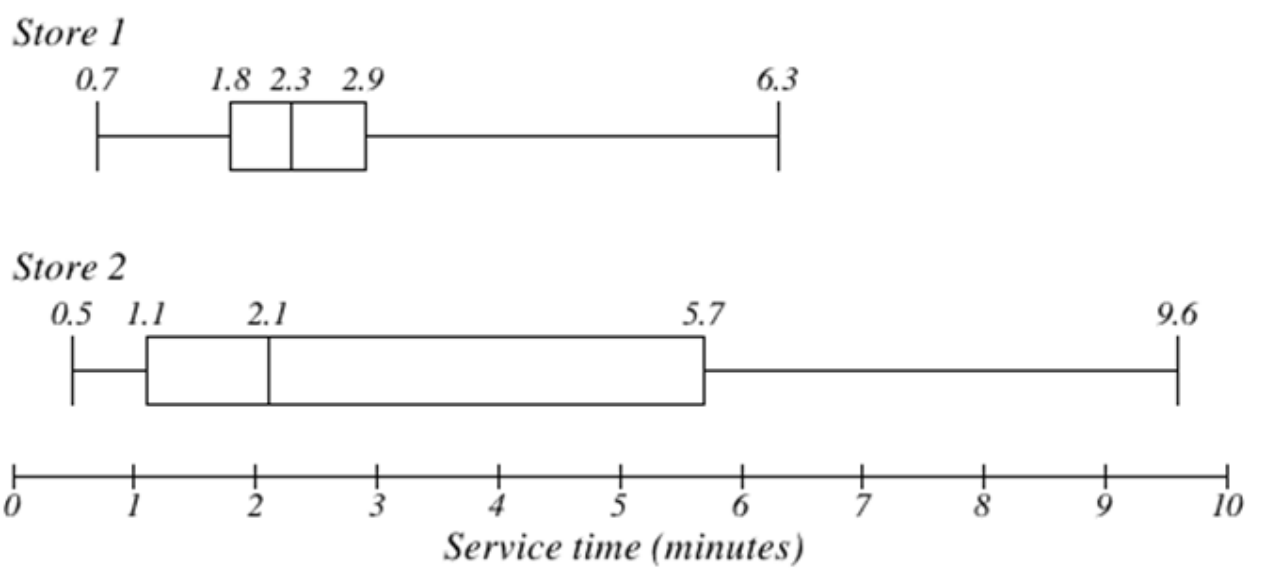
4.	If you want to let the calculator automatically check for outliers and draw the modified box plot if outliers exist, then select the first box plot icon for the **Type** in Step 2 (the one with the outliers.)	 
5.	If you press **[TRACE]** and use the left or right arrows, the calculator will show the values of the five-number summary, as well as the outlier. You can also create side-by-side box plots by entering data in multiple lists and turning more than one plot on at the same time.	

Comparing Multiple Box-and-Whisker Plots

Multiple box-and-whisker plots allow us to compare the overall features of several data sets quickly and efficiently. When two or more distributions are displayed side-by-side, differences in center, spread, shape, and outliers become immediately visible. This makes multiple box plots especially useful when the goal is to compare groups, examine trends, or highlight contrasts that may not be obvious from raw numbers alone. Placing box-and-whisker plots next to each other provides a clear visual comparison: the medians can be compared at a glance, the lengths of the boxes reveal any differences in the interquartile ranges, and the whiskers show how far each distribution extends. The next example illustrates how to interpret multiple box-and-whisker plots to draw meaningful conclusions about the similarities and differences among two data sets.

Example 3.4.6

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



While Store 2 had a slightly shorter median service time (2.1 minutes vs. 2.3 minutes), Store 2 shows less consistency in service times. That is, Store 2 has greater variation in the data.

At Store 1, 75% of customers were served within 2.9 minutes, while at Store 2, 75% of customers were served within 5.7 minutes.

Which store should you go to in a hurry? That depends upon your opinions about luck and probability – 25% of customers were served within 1.1 minutes at Store 2, while it took 1.8 minutes for 25% of the customers at Store 1 to be served. However, 25% of customers at Store 2 had to wait between 5.7 and 9.6 minutes for service, while all of the customers at Store 1 had been served within 6.3 minutes.

Example 3.4.7

The side-by-side box plots show salaries for actuaries and CPAs.



a. Kendra makes the median salary for an actuary. Kelsey makes the first-quartile salary for a CPA. Who makes more money, and by how much?

Solution

Kendra earns the actuary median salary of $90,000. Kelsey earns the CPA first-quartile salary of $40,000. Kendra earns $50,000 more than Kelsey.

b. Estimate the percentage of actuaries who make more than $95,000.

Solution

A salary of $95,000 corresponds to Q_3, the 75th percentile. Therefore, 25% of actuaries earn more than $95,000. (The top quartile always represents the highest 25% of the data.)

c. Which profession generally has the most variation in salary?

Solution

Although the overall ranges are similar, the middle 50% of CPA salaries are spread more widely. The CPA box -- which represents the IQR -- is longer than the actuary box. Therefore, CPA salaries show greater variation.

d. What percentage of actuaries make more than the median salary of a CPA?

Solution

The CPA median salary is about $75,000. This value aligns with Q_1 for actuaries. Since Q1 marks the 25th percentile, 75% of actuaries earn more than the CPA median salary.

e. Do you expect the mean or median actuary salary to be larger?

Solution

The actuary box plot is skewed left: the left whisker and left portion of the box are longer than those on the right. In a left-skewed distribution, the mean is pulled downward by lower values. Therefore, the median is larger than the mean for actuary salaries.

Section Summary

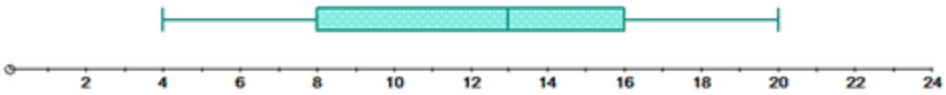
A box-and-whisker plot displays the five-number summary by showing a box that stretches from the lower quartile Q_1 to the upper quartile Q_3, with the median marked as a line inside the box. The whiskers extend from each quartile to the minimum and maximum non-outlier values. Because each whisker and each section of the box represents roughly 25% of the data, the lengths of these segments reveal how the data are spread within each quarter of the distribution. The width of the box itself is the interquartile range, which measures the spread of the middle 50% of the data.

Box-and-whisker plots are especially effective for summarizing the shape, center, and spread of a data set at a glance. When multiple box plots are placed on the same graph, they provide a clear visual comparison of these features across several groups.

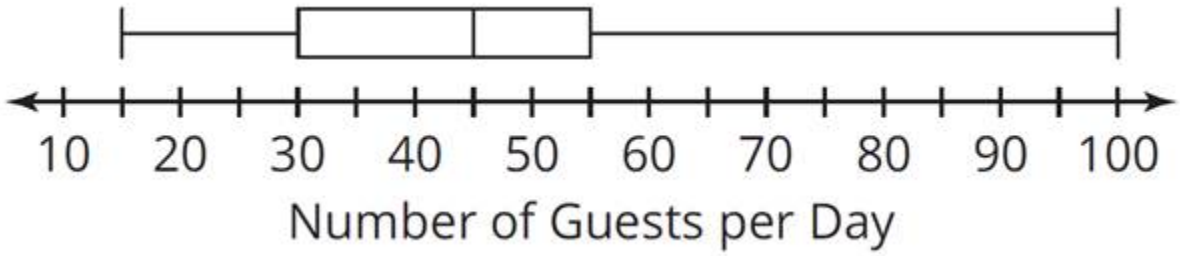
In this context, an outlier is any value that lies more than 1.5×IQR beyond either quartile. Outliers are shown as individual points rather than as part of a whisker, ensuring that unusually small or large values do not distort the visual impression of the main distribution.

Section 3.4 Review Questions

1. For the following box-and-whisker plot, list the five-number summary and comment on the shape of the distribution of the data.



2. The box-and-whisker plot below represents the number of rooms that were booked each night at an island resort during the latest tourist season.



   a. On what percent of the nights did the resort have 55 or fewer rooms booked?
   b. On what percent of the nights did the resort have 45 or more rooms booked?
   c. What is the IQR for the number of rooms booked at this resort?
   d. If this box-and-whisker plot shows data for the past 136 nights, on how many of the nights did the resort have 30 or fewer rooms booked?

3. Here are the data on the percentage of capacity for reservoirs in Idaho:

70	84	62	80	75	95
69	48	76	70	45	83
58	75	85	70	62	64
39	68	67	35	55	93
51	67	86	58	49	47
42	75				

   a. Find the five-number summary for this data set.
   b. Show all work to determine if there are outliers.
   c. Create a box-and-whisker plot (modified as necessary.)
   d. Describe the shape of the distribution of reservoir capacities in Idaho.
   e. Based on your answer in part (d), how would you expect the mean to compare to the median?
   f. Calculate the mean to verify your expectation in part (e).

4. Here are the data on the percentage of capacity of reservoirs in Utah.

80	46	83	75	83	90
90	72	77	4	83	105
63	87	73	84	0	70
65	96	89	78	99	104
83	81				

   a. Find the five-number summary for this data set.
   b. Show all work to determine if there are true outliers.
   c. Create a box-and-whisker plot (modified as necessary.)
   d. Describe the shape of the distribution of reservoir capacities in Utah.

e. Based on your answer in part (d), how would you expect the mean to compare to the median?
f. Calculate the mean to verify your expectation in part (e).

5. The following data represent the time (in minutes) that each of 20 people waited in line at a local bookstore to purchase the latest Harry Potter book:

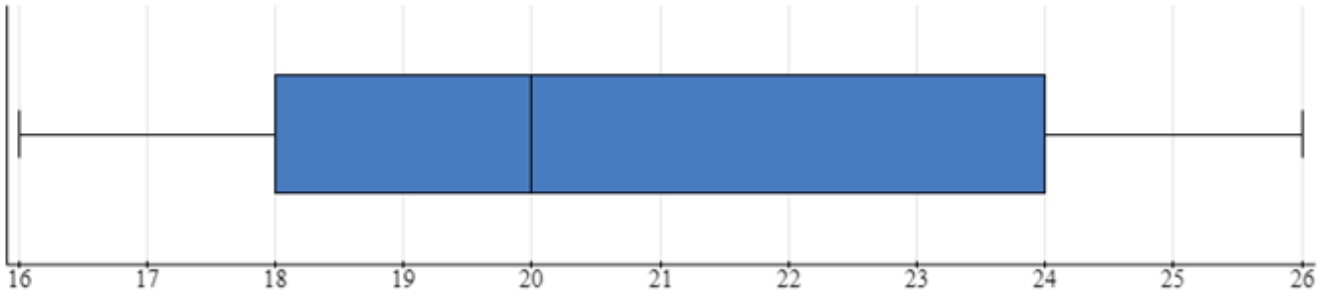
15	8	5	10	14
17	21	23	6	19
31	34	30	31	3
22	17	25	5	16

   a. Find the five-number summary for this data.
   b. Show all work to test for outliers.
   c. Graph the box-and-whisker plot for this data. Use a modified plot if necessary.

6. The lengths (in kilometers) of a random sample of 20 rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed below. Summarize the data with a box-and-whiskers plot. Use a modified plot if necessary.

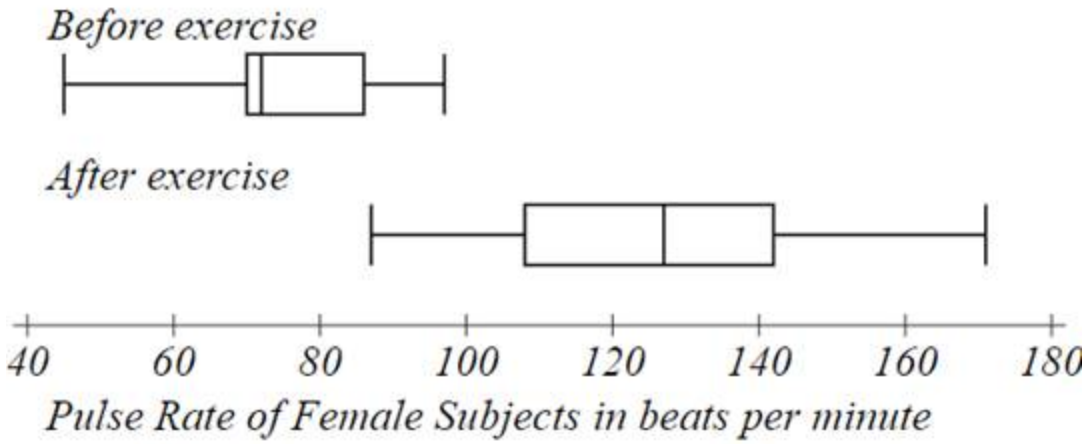
98	138	93	121	140
64	208	160	288	48
64	97	322	90	95
209	56	66	145	161

7. Given the box-and-whisker plot below, create a data set of 10 or more items that would generate the same plot.



8. If the median of a distribution is less than the mean, which of the following statements is the most correct?
   a. The distribution is skewed left.
   b. The distribution is skewed right.
   c. There are outliers on the left side.
   d. There are outliers on the right side.
   e. (a) or (c) could be true.
   f. (b) or (d) could be true.

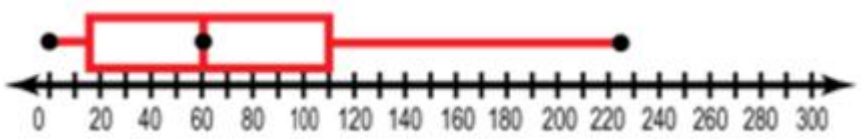
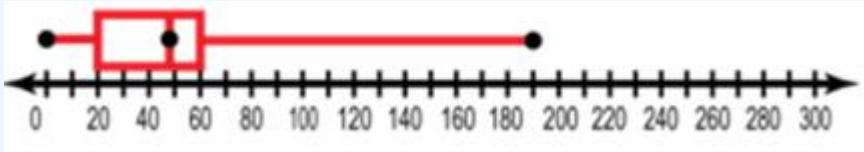
9. The box-and-whisker plot below shows pulse rates for a random sample of female subjects before and after one minute of exercise.



a. About 75% of female subjects had pulse rates after exercise of _____ or more.
b. Which group of pulse rates is more symmetric?
c. Which group has more consistent pulse rates in the middle 50% of the data?
d. Which group of pulse rates most likely has a larger standard deviation?
e. The typical pulse rate is higher for which group? Recall that 'typical' refers to a measure of center.

10. Use the graphs below to compare the amount of time a teenager spends in the bathroom getting ready to go to a party and the amount of time they spend in the bathroom getting ready for school. Answer the questions below.

Problem 10

Time Getting Ready for a Party (in minutes)	
Time Getting Ready for School (in minutes)	

a. What percent of teenagers spend at least 15 minutes getting ready for a party?
b. What is Q_3 for the time spent getting ready for school?
c. Is it more common for a teenager to spend more than 1 hour getting ready for school or more than 1 hour getting ready for a party? Explain.
d. Answer **True or False** for questions i through v.
   i. Some teenagers do not spend time getting ready for parties.
   ii. When getting ready for school, fewer students spend less than 20 minutes than for more than 60 minutes.
   iii. 25% of teenagers spend between 48 and 60 minutes getting ready for school.
   iv. A typical teenager spends more time getting ready for a party than getting ready for school.
   v. In general, teenagers are more consistent in the amount of time they spend getting ready for school than in the amount of time they spend getting ready for a party.

Answers

You Try It 3.4.A

a. $460
b. 50%
c. 75%
d. 25% of 80 = 20 students

This page titled 3.4: Box-and-Whisker Plots is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by Andy Jones.

- 11.6: Measures of Variation by David Lippman & Jeff Eldridge is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

3.5: Normal Distribution and Its Empirical Rule

Learning Objectives

- Describe characteristics of a normal distribution.
- Apply the *Empirical Rule* (68-95-99.7 Rule.)
- Compute a z-score for a value in a normal distribution.
- Compute an x-value in a normal distribution from its z-score.
- Use z-scores to compare relative standing.

Introduction

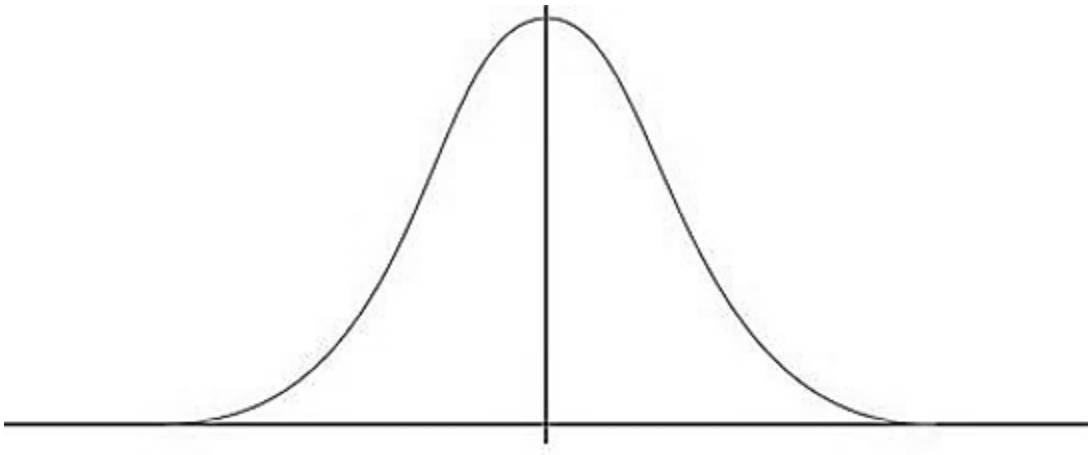
A normal distribution describes a pattern that appears again and again in real-world situations. You have likely seen this pattern without realizing it. When students enter a classroom between periods, a few arrive early, then more begin to show up, followed by the largest wave of arrivals, and finally the number tapers off as only a few students come in at the last moment. The same rising-and-falling pattern occurs when popcorn kernels begin to pop or when you measure characteristics such as height, shoe size, or hand width in a group of children: a few measurements are very small, a few are very large, and most cluster around a central value.

These examples illustrate a common shape in statistics -- one that increases gradually, reaches a peak, and then decreases in a similar way. Because this pattern is so widespread, statisticians call it the **normal distribution**. The normal distribution is one of the most important models in statistics. Many measurements in nature, psychology, education, and everyday life closely follow this pattern, and many theoretical results in statistics are built upon it. Although real data sets are never *perfectly* normal, many come close enough that the idealized normal curve becomes a powerful tool for understanding and predicting behavior in practical situations.

In this section, we explore the fundamental properties of the normal distribution and learn how this model helps us interpret data that follow a similar pattern.

Characteristics of a Normal Distribution

A **normal distribution** has distinctive and consistent features in its shape, center, and spread. Understanding these features allows us to recognize when real data follow a pattern that is approximately normal and to use the normal model effectively.

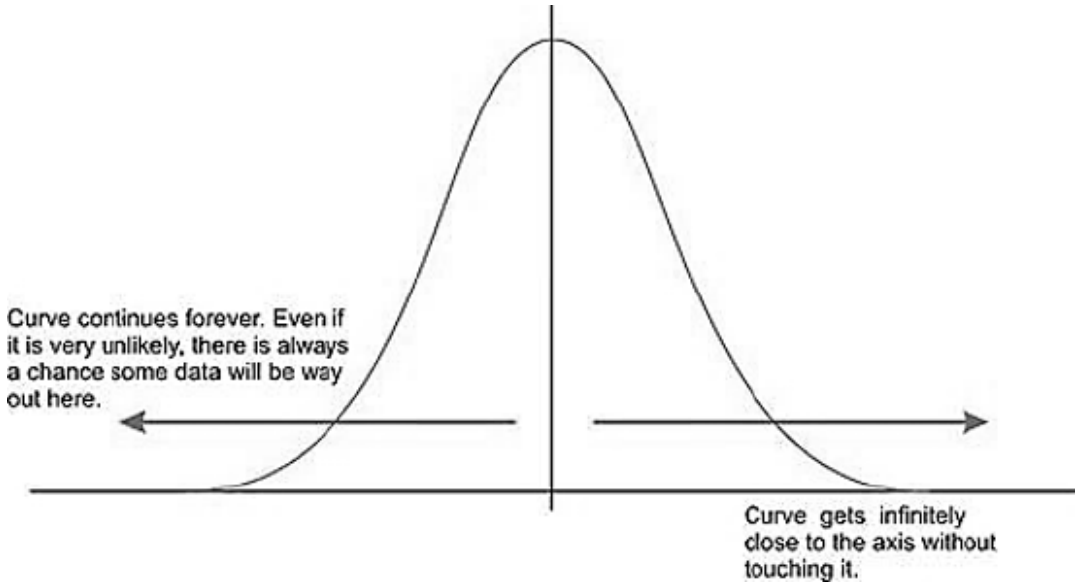


The Normal Distribution

Shape

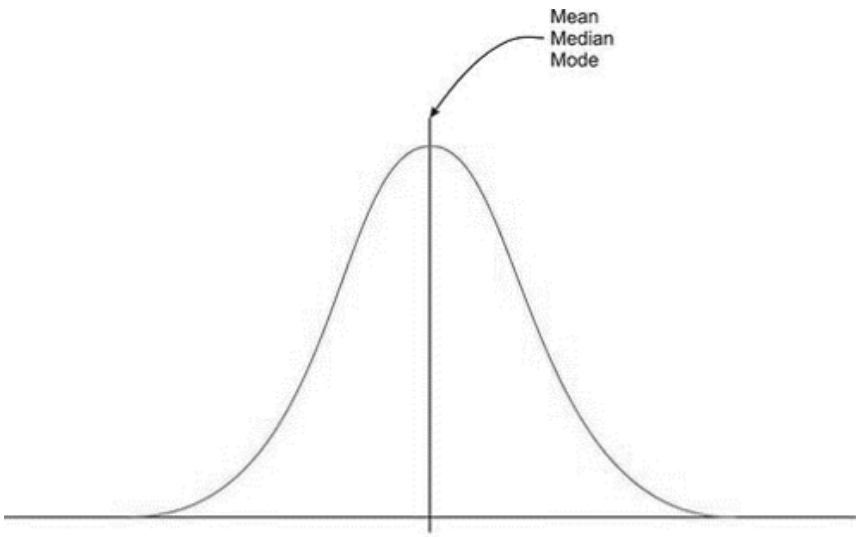
A normal distribution is mound-shaped, symmetric, and unimodal. Because of its smooth, balanced form, it is often called the *normal curve* or *bell curve*. In its idealized mathematical form, the curve extends indefinitely in both directions, approaching but never touching the x-axis. This asymptotic behavior reflects the idea that extremely small or extremely large values are possible but occur with very low frequency.

The symmetry of the curve means that the left and right halves are mirror images. Values become less frequent as they move away from the center, but the rate of decrease is the same in both directions.

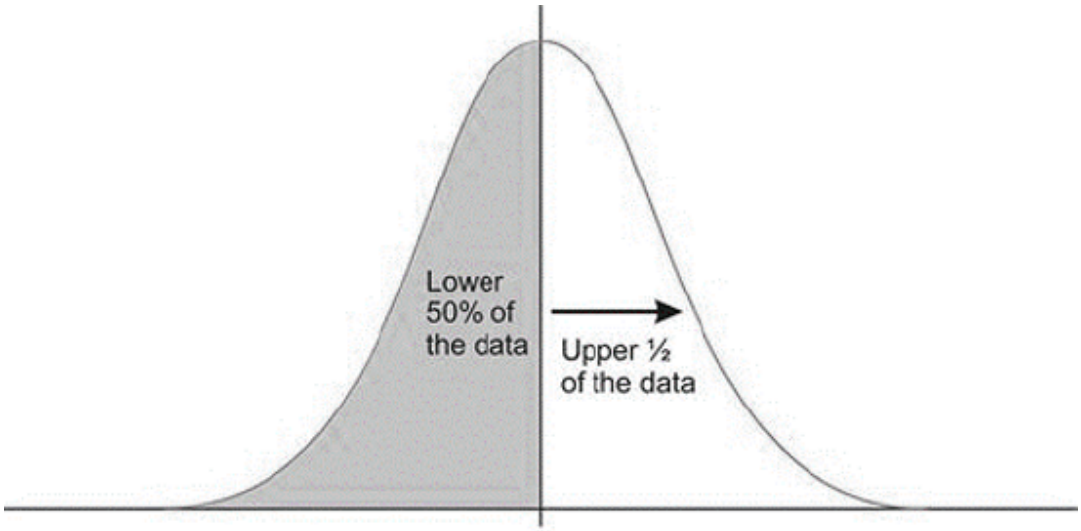


Center

In a perfectly normal distribution, all common measures of center -- the mean, median, and mode -- coincide at the same point. This central location lies directly beneath the highest point of the curve and divides the distribution into two equal halves. It corresponds to the 50th percentile, meaning half the data fall below this value and half above it. Because we tend use the mean when a distribution is symmetric, we use the Greek letter μ (pronounced "*mew*") for the mean when working with a normal distribution.



Normal Distribution Center



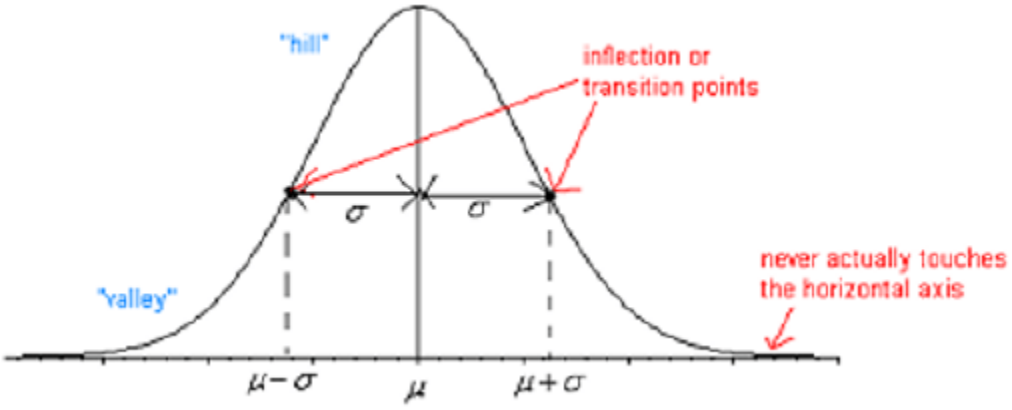
Normal Distribution divided into equal halves by the center

Spread

Because a normal curve extends infinitely, the range is not a meaningful measure of spread. Instead, the spread of a normal distribution is described using the standard deviation, which measures the typical distance of data values from the mean. We use the Greek letter σ for the standard deviation when working with a normal distribution.

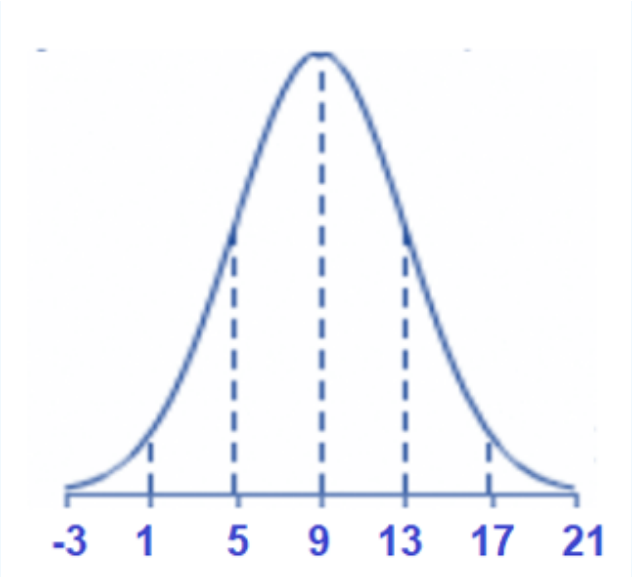
Normal curves with small standard deviations are tall and narrow, indicating that most values cluster tightly around the mean. Curves with larger standard deviations are wider and flatter, showing that values are more dispersed and that the tails contain more observations.

A useful visual feature of the normal curve is its inflection points—the points where the curve changes from curving upward to curving downward. These transition points occur exactly one standard deviation from the mean. Thus, the distance from the mean to either inflection point provides a direct visual estimate of σ.



Example 3.5.1

Use the graph of the normal distribution to identify the values of μ and σ.



Solution

The mean μ is located at the value directly below the peak of the curve, so $\mu = 9$. The inflection points occur at 5 and 13, each 4 units away from the mean. Therefore, the standard deviation is $\sigma = 4$. Standard deviation represents distance and is always a non-negative quantity.

The Empirical Rule

For any distribution that is approximately normal, a remarkable and consistent pattern emerges in how data are spread around the mean. This pattern is known as the ***Empirical Rule***, or the **68–95–99.7 Rule**, and it provides quick approximations of the proportion of observations falling within one, two, or three standard deviations of the mean. Although more precise methods for finding relative frequencies will be introduced in the next section, the *Empirical Rule* offers an efficient way to estimate probabilities and understand the behavior of normal data.

Empirical Rule for Normal Distributions

Because a normal distribution is symmetric and centered at its mean μ, the distance from the mean is measured using the standard deviation σ.

The *Empirical Rule* states

\(P(μ–σ

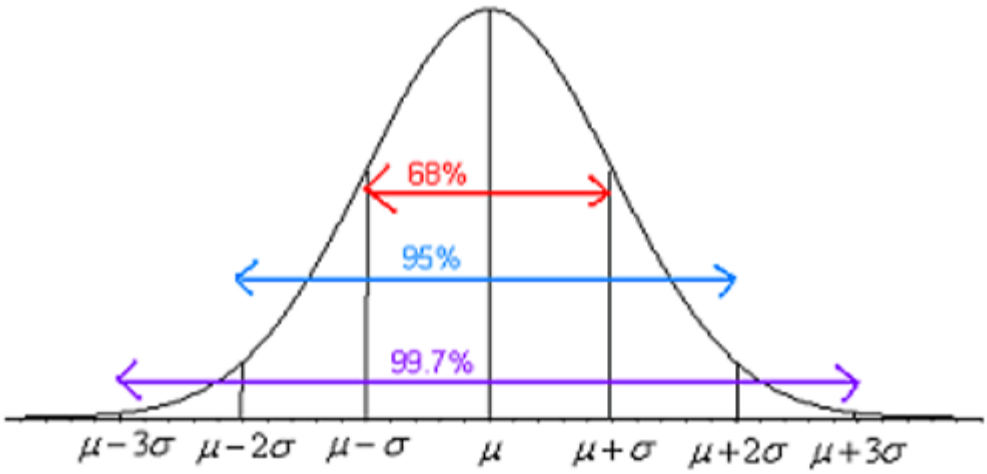
Approximately 68% of the data lie within one standard deviation of the mean.

\(P(μ–2σ

Approximately 95% of the data lie within two standard deviations of the mean.

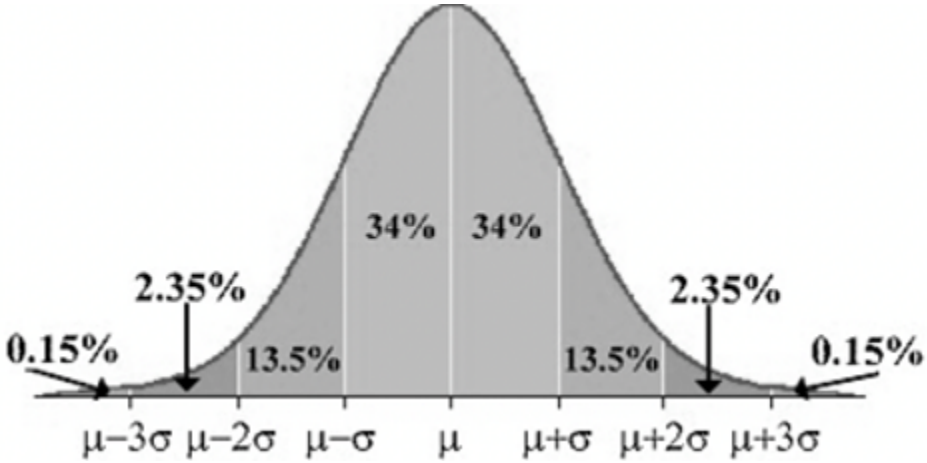
\(P(μ–3σ

Approximately 99.7% of the data lie within three standard deviations of the mean.



These percentages do not account for *all* possible values. A small amount of area remains in the extreme tails beyond $\mu \pm 3\sigma$. Since the total area under a normal curve is 100%, subtracting 99.7% leaves 0.3% of the data in both tails combined. By symmetry, each tail contains 0.15% of the distribution.

When we divide the normal curve into eight intervals -- at each of the three standard deviation positions on each side of the mean -- we obtain a more detailed picture of how data accumulate across the distribution. These intervals help illustrate where data are most concentrated and how quickly frequencies diminish as values move farther from the mean.



The *Empirical Rule* is most effective when the endpoints of an interval fall exactly 1, 2, or 3 standard deviations from the mean. When the boundaries are not integer multiples of σ, we must use technology or probability tables, as demonstrated in Section 3.6.

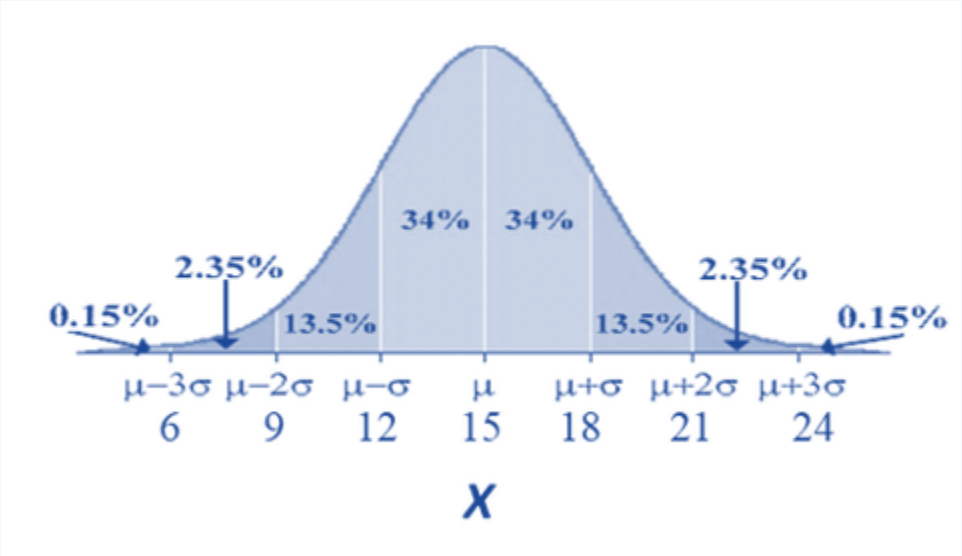
Example 3.5.2

The mean age at which children begin walking is 15 months, and the standard deviation is 3 months. Assume *walking age* is normally distributed.

a. Draw a sketch of the distribution for *walking age*. Label the axis to three standard deviations on each side of the mean.

Solution

Begin by marking the mean walking age of 15 months directly beneath the peak of the normal curve. From this center point, move left in increments of one standard deviation: subtract 3 months once to locate $\mu - \sigma = 12$, subtract again to locate $\mu - 2\sigma = 9$, and subtract a third time to locate $\mu - 3\sigma = 6$. Repeat the process on the right side of the mean: add 3 months to locate $\mu + \sigma = 18$, add again for $\mu + 2\sigma = 21$, and add once more for $\mu + \sigma = 24$. After the axis is labeled, use the *Empirical Rule* to mark the percentage of area in each region of the curve. These labeled intervals provide a visual guide for estimating percentages associated with walking ages.



b. What percentage of children start walking between 12 and 18 months of age?

Solution

In the sketch from part a, the interval from 12 to 18 months corresponds to $\mu - \sigma = 15 - 3 = 12$ and $\mu + \sigma = 15 + 3 = 18$. The *Empirical Rule* tells us that approximately 68% of all observations fall within one standard deviation of the mean.

\(P(12

Thus, about 68% of children begin walking between 12 and 18 months.

c. What percentage of children walk for the first time before age 18 months?

Solution

To find $P(X < 18)$, add the percentages in all regions to the left of $\mu + \sigma$. Using the detailed *Empirical Rule* breakdown:

- Below $\mu - 3\sigma$: 0.15%
- Between $\mu - 3\sigma$ and $\mu - 2\sigma$: 2.35%
- Between $\mu - 2\sigma$ and $\mu - \sigma$: 13.5%
- Between $\mu - \sigma$ and μ: 34%
- Between μ and $\mu + \sigma$: 34%

Adding these gives $P(X < 18) = 0.15\% + 2.35\% + 13.5\% + 34\% + 34\% = 84\%$.

Alternatively, you could add the 50% below the mean to the 34% between μ and $\mu + \sigma$: $50\% + 34\% = 84\%$.

Either way, approximately 84% of children walk before reaching 18 months.

d. Approximate the percentage of children who start walking after 21 months of age.

Solution

To find $P(X > 21)$, add the percentages in the two right-tail regions beyond $\mu + 2\sigma$:

- Between $\mu + 2\sigma$ and $\mu + 3\sigma$: 2.35%
- Beyond $\mu + 3\sigma$: 0.15%

$P(X > 21) = 2.35\% + 0.15\% = 2.5\%$

Thus, about 2.5% of children begin walking after 21 months.

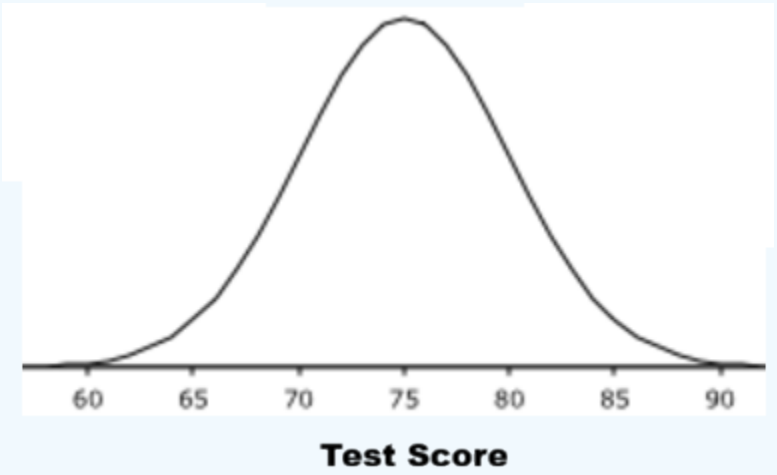
Example 3.5.3

The mean score on a standardized test is 75 points with standard deviation 5 points. The scores follow an approximately normal distribution. Sketch a normal curve and use it to answer these questions.

a. What percentage of the students had scores between 65 and 85?
b. What percentage of the students had scores between 65 and 75?
c. What percentage of the students had scores between 70 and 80?
d. What percentage of the students had scores above 85?

Solution

To solve each of these questions, it would be helpful to draw the normal curve that represents this situation. The mean is 75, so place 75 below the peak. The standard deviation is 5, so to label the axis, add 5 to the mean 3 times and subtract 5 from the mean 3 times. The graph looks like the following:



a. From the graph, $X = 65$ and $X = 85$ are both two standard deviations from the mean. According to the *Empirical Rule*, this percentage must be 95%.
b. The scores from $X = 65$ to $X = 75$ make up half of the percentage from $X = 65$ to $X = 85$. Because of symmetry, that means that $P(65 < X < 75$ half of the answer to part a: $\frac{1}{2} \times 95\% = 47.5\%$. You could have also added $13.5\% + 34\%$.
c. From the graph, $X = 70$ and $X = 80$ are both one standard deviation from the mean. According to the *Empirical Rule*, this percentage must be 68%.
d. One way to find $P(X > 85)$ is to add $2.35\% + 0.015\% = 2.5\%$. There are other strategies too.

You Try It 3.5.A

The distribution of wait times at a drive-through restaurant is approximately normal with mean 185 seconds and standard deviation 15 seconds. Use a sketch and the *Empirical Rule* to find the percentage of wait times that are

a. between 155 and 215 seconds.
b. more than 170 seconds.
c. less than 155 seconds.
d. between 170 and 215 seconds.

Answer available at the end of this section.

z-Scores

In Section 2.3, we defined a **z-score** for sample data as the number of standard deviations an observed value x lies from the sample mean $\bar{x}$. In the context of a normal distribution, the idea is the same, but now the z-score measures how far a value lies from the population mean μ. A z-score tells us the position of a value relative to the entire distribution, expressed in units of the standard deviation σ.

A positive z-score indicates a value above the mean, a negative z-score indicates a value below the mean, and a z-score of zero corresponds exactly to the mean. Because normal distributions are symmetric, z-scores provide a standardized way to compare values, interpret relative standing, and connect raw data to the *Empirical Rule*.

The formula for a z-score of any value x when compared to a normal distribution is

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

This calculation converts any value x into a standardized measurement that tells how many standard deviations it is from the mean.

As saw in Section 2.3, the z-score formula can be rearranged to solve for x. This form is especially useful when you know the z-score, the mean, and the standard deviation.

$$x = \mu + z\sigma$$

Working with z-Scores on the Normal Curve

Suppose scores on a math test are normally distributed with mean $\mu = 82$ and standard deviation $\sigma = 7$. Note that sometimes you may see this written in an abbreviated way as $N(82, 7)$.

- A score of 89 is 7 points *above* the mean. Since 7 equals one standard deviation, the z-score is $z = +1$.
- A score of 75 is 7 points *below* the mean. This is one standard deviation below the mean, so $z = -1$.
- If a student's z-score is -2, the student scored exactly two standard deviations below the mean: $x = 82 - 2(7) = 68$.

Because the distribution is normal, z-scores also tell us how a score compares to the rest of the class. From the *Empirical Rule*, about 68% of students scored between $z = -1$ and $z = 1$, or between 75 and 89. About 16% scored above 89, and 16% scored below 75.

Example 3.5.4

Suppose the heights of college basketball players are $N(76, 2)$ inches.

a. Darren is 79 inches tall. Find his z-score.

Solution

$$z = \frac{79 - 76}{2} = \frac{3}{2} = 1.5$$

Darren scored 1.5 standard deviations above the mean.

Solution

b. Marcus is 73.5 inches tall. Find his z-score.

$$z = \frac{73.5 - 76}{2} = \frac{-2.5}{2} = -1.25$$

Marcus is 1.25 standard deviations below the mean.

Solution

c. Hector has a z-score of 1.75. How tall is he?

$$x = \mu + z(\sigma) = 76 + 1.75(2) = 79.5$$

Hector is 79.5 inches tall, which is 3.5 inches above the mean -- consistent with a z-score of 1.75.

You Try It $3.5.B$

The daily protein intake for adult endurance athletes is normally distributed as $N(110,\ 18)$ grams.

a. Juan consumes 140 grams of protein per day. Find his z-score.

b. Ava has a z-score of -0.75. How much protein does she consume per day?

Answer available at the end of this section.

Comparing with z-Scores

One of the most powerful uses of z-scores is comparing observations from different distributions. When raw scores come from groups with different means and/or standard deviations, converting them to z-scores places them on the same scale. This process is called **standardizing** the distribution.

Example $3.5.5$

Resting heart rate among healthy adults is approximately normally distributed with mean $\mu = 72$ beats per minute (bpm) and standard deviation $\sigma = 8$ bpm.

Daily step counts among adults in a large city are also approximately normally distributed with mean $\mu = 6,800$ steps and standard deviation $\sigma = 1,900$ steps.

James has a resting heart rate of 60 bpm. Maryam walks 9,200 steps per day.

Who is more unusual within their own group -- James or Maryam?

Solution

Raw values alone cannot tell us how unusual each measurement is. To compare fairly, convert each value to a z-score, which expresses distance from the mean in units of standard deviation.

For James, $z = \frac{60-72}{8} = \frac{-12}{8} = -1.5$. James's resting heart rate is 1.5 standard deviations below the population mean.

For Maryam, $z = \frac{9{,}200-6{,}800}{1{,}900} = \frac{2{,}400}{1{,}900} \approx 1.26$. Maryam's daily step count is 1.26 standard deviations above the population mean.

To determine who is more unusual relative to their population, compare the magnitudes of the z-scores:

$$|z_{James}| = 1.5$$

$$|z_{Maryam}| \approx 1.26$$

Because $1.5 > 1.26$, James's resting heart rate is farther from the average of his population than is Maryam's step count is from her population average.

You Try It $3.5.C$

To prepare for applying for college, Jessica took the SAT and Cory took the ACT.

- Jessica scored 1120, where SAT scores are $N(1000, 200)$.
- Cory scored 24, where ACT scores are $N(21, 5)$.

Use z-scores to compare their performances.

Answer available at the end of this section.

Section Summary

A normal distribution is an idealized model used to describe a variable in which values near the center occur most frequently and values farther from the center occur less often. Its shape is perfectly symmetric, smooth, and unimodal, forming the familiar bell-shaped curve. In a normal distribution, the mean, median, and mode all coincide at the highest point of the curve, and the curve extends indefinitely in both directions. The standard deviation determines the spread of the distribution. On the graph, σ is the distance from the mean to each of the curve's inflection points, where the curvature changes.

The *Empirical Rule* summarizes how data are distributed around the mean in a normal model. Approximately 68% of all values fall within one standard deviation of the mean, 95% fall within two standard deviations, and 99.7% fall within three. These proportions allow quick estimates of how typical or unusual a value is without using technology or tables.

A z-score provides a standardized measure of position within a normal distribution. It tells how many standard deviations a value x lies from the mean μ, using the formula

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

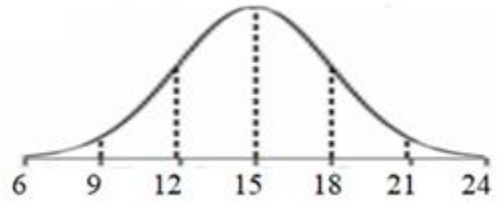
When an observation's z-score is known for an observation in a normal distribution, the value of the observation can be found by rearranging the z-score formula:

$$x = \mu + z\sigma.$$

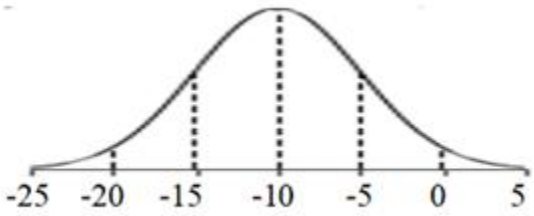
A positive z-score indicates a value above the mean, a negative z-score indicates a value below the mean, and a z-score of zero corresponds exactly to the mean. As the absolute value $|z|$ increases, the value becomes more unusual relative to the rest of the distribution. Because z-scores place all values on the same scale, they allow meaningful comparisons across different normal distributions with different centers and spreads.

Section 3.5 Review Questions

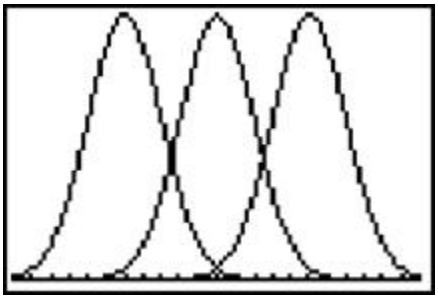
1. Identify the values of μ and σ using the graph of the normal distribution shown below.



2. Identify the values of μ and σ using the graph of the normal distribution shown below.

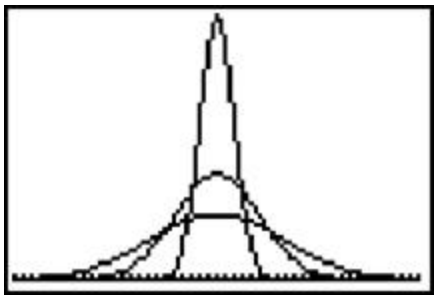


3. The sketch has 3 normal curves plotted on the same set of axes.



a. Compare the means of the 3 distributions.
b. Compare the standard deviations of the 3 distributions.

4. The sketch has 3 normal curves plotted on the same set of axes.



a. Compare the means of the 3 distributions.
b. Compare the standard deviations of the 3 distributions.

5. The distribution of wait times at a drive-through restaurant is approximately normal with mean 185 seconds and standard deviation 15 seconds.
   a. Draw a well-labeled sketch of this normal model.
   b. Use the *Empirical Rule* to find the percentage of wait times that are
      i. more than 170 seconds.
      ii. less than 155 seconds.
      iii. between 170 and 215 seconds

6. The weight of 1-year-old girls in the US is normally distributed with mean of 9.5 kilograms and standard deviation of 1.1 kilograms.
   a. Draw a well-labeled sketch of this normal model.
   b. Use the *Empirical Rule* to find the percentage of these 1-year-old girls who weigh
      i. less than 8.4 kg
      ii. between 7.3 kg and 11.7 kg
      iii. more than 12.8 kg

7. The mean mathematics SAT score in 2018 was 514 with a standard deviation of 117. Assume mathematics SAT scores are normally distributed. Draw a well-labeled sketch of a normal model, shade the proper region, and use the *Empirical Rule* to estimate the proportion of SAT test takers who meet the following conditions:
   a. score between 397 and 748.
   b. score at least 631.

8. A dishwasher has a mean life of 12 years with standard deviation of 1.25 years. Assume the life of a dishwasher is normally distributed.
   a. How do you know that you can the *Empirical Rule* to approximate percentages of dishwashers that have certain lifetimes?
   b. Use the *Empirical Rule* to approximate the percentage of dishwashers lasting between 9.5 years and 13.25 years.
   c. The warranty guarantees the dishwasher will last at least 9.5 years or your money back. On what percentage of the dishwashers sold will the company have to refund money to a customer?

9. Which *one* of these variables is most likely to be normally distributed? For the other choices, explain why you believe they would not follow a normal distribution.
   a. The hand span (measured from the tip of the thumb to the tip of the extended 5th finger) of a random sample of high school seniors.
   b. The annual salaries of all employees of a large shipping company.
   c. The annual salaries of a random sample of 50 CEOs of major companies, 25 women and 25 men.
   d. The dates of 100 pennies taken from a cash drawer in a convenience store.

10. The scores on a standardized reading test for a group of 4th graders $N(81, 6.25)$.
    a. Calculate the z-score for each of the following exam scores: 65, 81, and 93.
    b. Student X has a z-score of 1.45. What raw score did Student X earn?
    c. Student Y has a z-score of -2.40. What raw score did Student Y earn?

11. Classes A and B have final exam scores that are approximately normally distributed, with the mean for Class A equal to 72 and the mean for Class B equal to 82. The standard deviation of Class A's scores is 10, and the standard deviation of Class

B's scores is 5.

a. For which class is a score of 90 more impressive? Support your answer with z-scores.
b. For which class a score of 60 more discouraging? Support your answer with z-scores.

12. A runner for Team 1 runs a race in 57 seconds. Team 1 has running times with a mean of 64.2 seconds and a standard deviation of 4.2 seconds. A runner for Team 2 can run a race in 58 seconds. Team 2 has running times with a mean of 62.1 seconds and a standard deviation of 1.2 seconds.
    a. Which runner is faster? (This is not a trick question...no statistics involved here.)
    b. Which runner is faster relative to their peers? (Now you have to use statistics.)

Answers

You Try It $3.5.A$

a. 95%
b. 84%
c. 2.5%
d. 81.5%

You Try It $3.5.B$

a. $z = 1.67$
b. 96.5 grams

You Try It $3.5.C$

Jessica and Cory performed equally well. Their z-score are both 0.6, showing show the same level of performance even though the raw scores are on different scales.

This page titled 3.5: Normal Distribution and Its Empirical Rule is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

- 2.4: The Normal Distribution 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.

3.6: Applying a Normal Distribution

Learning Objectives

- Find percentages under a normal distribution using technology.
- Find normal-distribution values given various percentages.
- Find quartiles and percentiles of a normal distribution using technology.
- Solve real-life normal model problems using technology.

Introduction

In Section 3.5, the *Empirical Rule* allowed us to estimate relative frequencies for normally distributed variables, though only for intervals located exactly 1σ, 2σ, and 3σ from the mean. While these benchmarks provide quick approximations, they apply to a limited set of fixed distances.

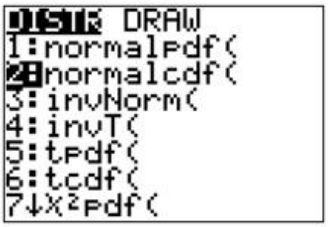
In this section, we continue working with the normal distribution but extend our methods beyond those fixed intervals. Our focus now shifts to questions requiring greater precision for problems involving intervals for which the *Empirical Rule* cannot supply an answer.

Finding Normal Percentages Using a Calculator

Before modern calculators and statistical software were widely available, finding percentages under the normal curve required computing z-scores and consulting printed tables of normal areas. Today, technology allows us to compute these percentages directly, and in this course we will assume the use of a calculator or computer for these normal distribution problems. Even so, sketching the normal curve and identifying the interval of interest remains an essential step in understanding the problem.

The table below outlines the general procedure for computing normal percentages on the TI-83/84 calculator using the **normalcdf(** command, found in the Distribution menu.

Computing Percentages Under the Normal Curve on the TI-83/84 Calculator

1.	Push **[2nd] [VARS]** to open the **DISTR** (Distribution) menu. Choose **2:normalcdf(**.	

2.	You will have one of the two screens shown here: If you have a screen like the one on the top with a blinking cursor, you need to type the lower endpoint of the interval, the upper endpoint of the interval, the mean, and the standard deviation. Then, close the parentheses. If you have a STAT WIZARD screen that looks like the one on the bottom, enter the same information in the table. NOTE: If the lower endpoint is -∞, enter -10^{99} by pushing **[(-)] 10 [^] 99**. If the upper endpoint is ∞, enter 10^{99} by pushing **10 [^] 99**. These are close enough approximations for the infinities when we use them in statistics.	
3.	Press **[ENTER]**, or highlight **Paste** and push **[ENTER]**.	

Example 3.6.1

Lengths of human pregnancies follow a normal distribution with mean $\mu = 272$ days and standard deviation $\sigma = 9$ days.

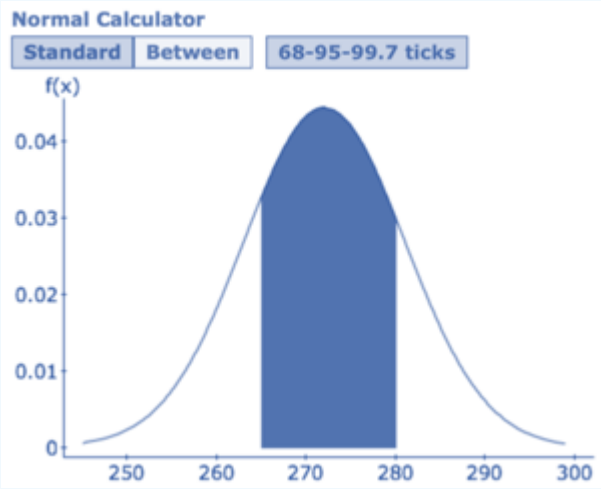
a. Find the percentage of pregnancies that last between 265 and 280 days.

Solution

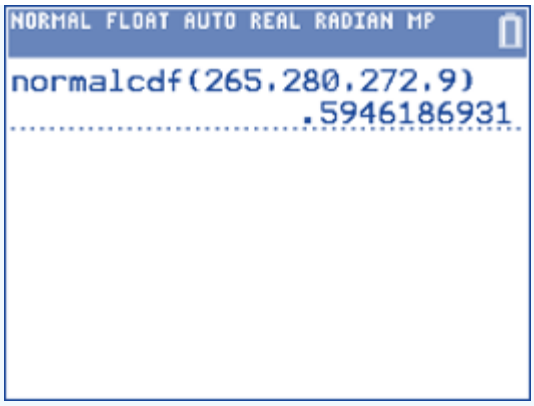
Translate the question into a mathematical statement:

\(P(265

Sketch a normal curve centered at 272 with standard deviation 9. We shade the portion of the curve we are calculating: between 265 and 280.



Use the **normalcdf(265, 280, 272, 9)** command. The lower endpoint of the interval is 265 and the upper endpoint of the interval is 280.



\(P(265

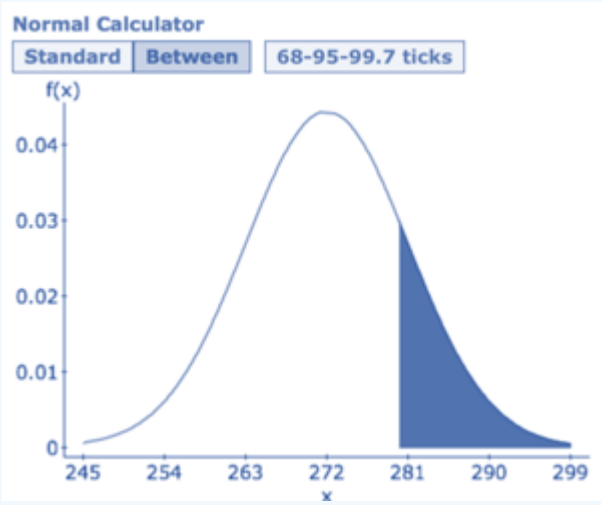
b. Find the percentage of pregnancies that last more than 280 days.

Solution

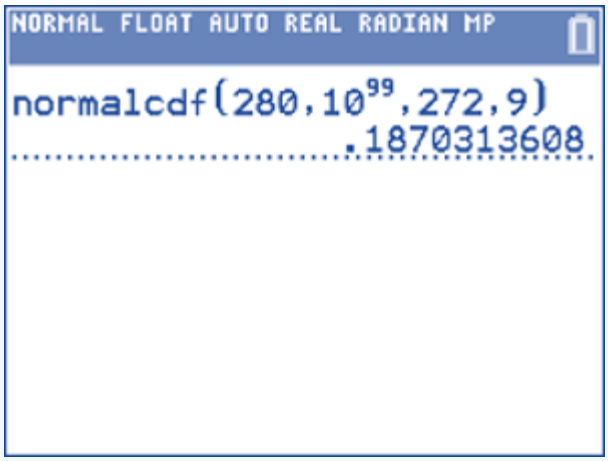
Translate the question into a mathematical statement:

$P(X > 280)$.

Sketch a normal curve centered at 272 with standard deviation 9. We shade the portion of the curve we are calculating: to the right of 280.



Use the **normalcdf(280, ∞, 272, 9)** command. The lower endpoint of the interval is 280 and the upper endpoint of the interval is +∞.



$P(X > 280) \approx 0.1870$. Approximately 18.70% of pregnancies last more than 280 days.

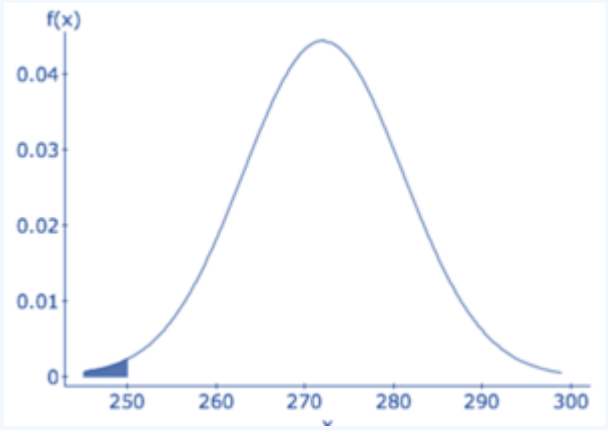
c. Find the percentage of pregnancies that last less than 250 days.

Solution

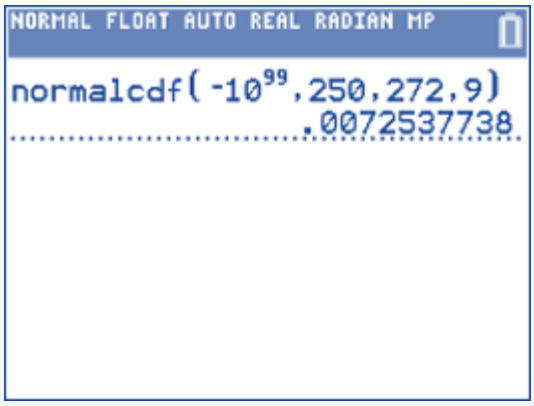
Translate the question into a mathematical statement:

$P(X < 250)$.

Sketch a normal curve centered at 272 with standard deviation 9. We shade the portion of the curve we are calculating: to the left of 250.



Use the **normalcdf(−∞, 250, 272, 9)** command. The lower endpoint of the interval is -∞ and the upper endpoint of the interval is 250.



$P(x < 250) \approx 0.0073$. About 0.73% of pregnancies last fewer than 250 days -- an unusually small proportion.

You Try It 3.6.A

Daily water intake among long-distance runners is approximately normally distributed with mean $\mu = 2.8$ liters and standard deviation $\sigma = 0.45$ liters.

a. What percentage (to the nearest whole percent) of runners drink between 3.0 and 3.6 liters per day?
b. What percentage (to the nearest whole percent) of runners drink more than 2.5 liters of water per day?

Answer available at the end of this section.

Solving Unknown Value Problems

In Example 3.6.1, we determined the relative frequency of pregnancies falling within specified intervals of the variable x. In each case, whether $265 < x < 280$, $x > 280$, or $x < 250$, we were given an interval and asked to find the corresponding percentage. To do this, we used the normalcdf(command on the TI-83/84 calculator.

We now turn to the reverse problem: instead of finding a percentage from a given value of x, we will find the value of x that corresponds to a specified percentage. In other words, we seek the point on the horizontal axis that "cuts off" a given proportion of the distribution. This type of problem is common in statistical applications, and technology makes the computation straightforward. On the TI-83/84 calculator, the relevant command is **invNorm(**, found in the Distribution menu.

Example 3.6.1

A normal distribution has mean $\mu = 35$ and standard deviation $\sigma = 7.4$. Find the value x such that 80% of the curve lies left of x.

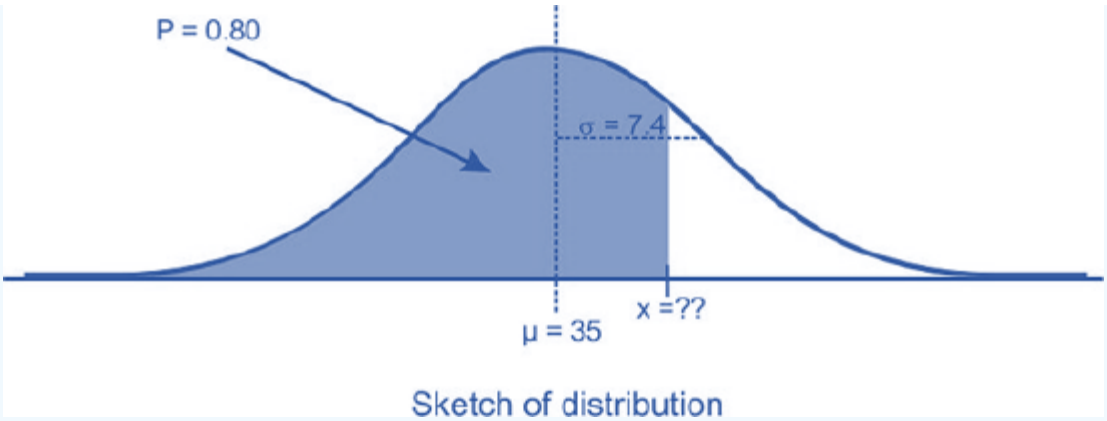
Solution

In this problem, the percentage is known and the cutoff value x is unknown.

Translate the question into a mathematical statement:

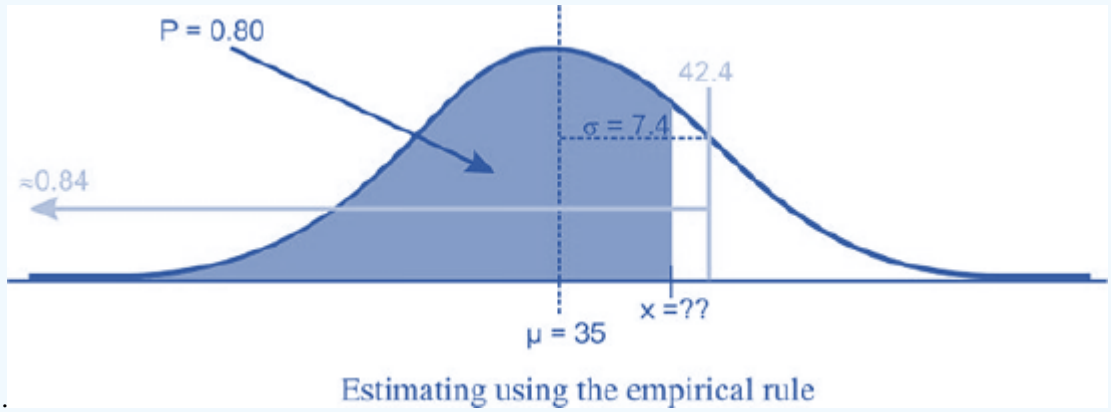
$P(X < \ ?) = 0.80$.

Sketch a normal curve centered at 35 with standard deviation 7.4. We shade the left portion so that it contains 80% of the curve.



Sketch of distribution

If we had to *estimate* an answer, we know from the *Empirical Rule* that about 84% of the area lies left of $\mu + \sigma$ as the sketch below shows. Therefore, we expect the value of x that cuts off the lowest 80% of the curve to be slightly less than $35 + 7.4 = 42.4$.

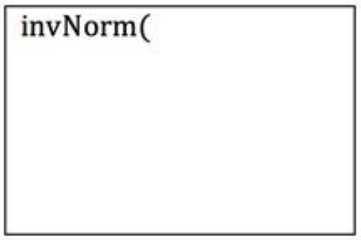
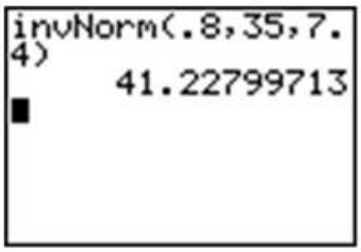


Estimating using the empirical rule

While 42.4 is a reasonable estimate for x, we would like to have a more precise answer. The following instructions show how to use a TI calculator to find a more precise value for x so that $P(X < \ ?)$ is 80%.

Computing x-values of Normal Distributions on the TI-83/84 Calculator

1.	Push **[2nd] [VARS]** to open the **DISTR** (Distribution) menu. Choose **3:invNorm(.**	DISTR DRAW 1:normalpdf(2:normalcdf(3:invNorm(4:invT(5:tpdf(6:tcdf(7↓χ²pdf(

2.	You will have one of the two screens shown: If you have a screen like the one on the top with a blinking cursor, enter the ***area to the left**** of the x-value , the mean, and the standard deviation – all separated with commas. Close the parentheses. If you have a STAT WIZARD screen that looks like the one on the bottom, enter the same information in the table.	 
3.	Press **[ENTER]**, or highlight **Paste** and push **[ENTER]**.	
	*NOTE: On most TI calculators, you will need to give the left-sided area. Depending on the problem, you may know a right-sided area. If this is the case, you will need to subtract from 1 to get the left-sided area. For example, if the area to the right of x is 0.1, then the area to the left of x would be $1 - 10.1 = 0.9$ and 0.9 is the first number entered after the **invNorm(** command on the calculator.	

As the calculator shows, the value of $x \approx 41.23$ splits the normal distribution curve into two parts so that the percentage $P(x < 41.23) \approx 80\%$ and $P(x > 41.23) \approx 20\%$. This is close to our estimate of 42.4 using the *Empirical Rule*.

Example 3.6.3

Recall in Example 3.6.1 that lengths of human pregnancies are $N(272, 9)$. Find the lengths that make up the shortest 10% of all human pregnancies, rounded to the nearest whole day.

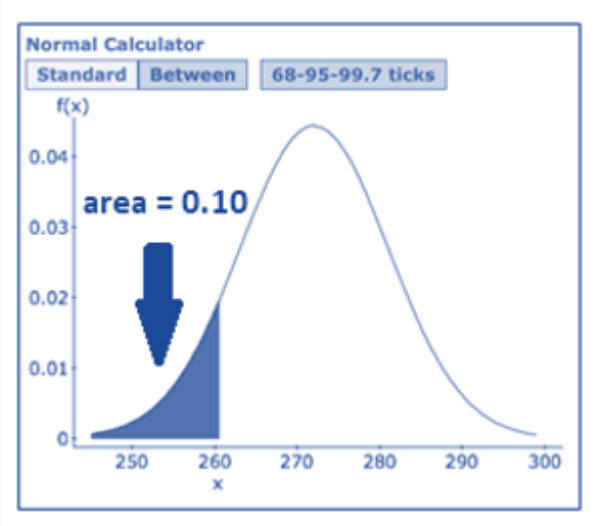
Solution

This problem asks for an unknown cutoff value on the horizontal axis such that 10% of pregnancies are to its left.

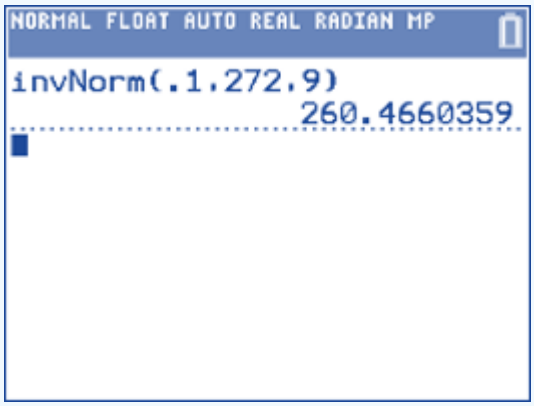
Translate the question into a mathematical statement:

$$P(X < \ ?) = 0.10.$$

Sketch a normal curve centered at 272 with standard deviation 9. We shade the left portion so that it contains 10% of the curve.



Use the **InvNorm(0.10, 272, 9)** command. The area to the left of the unknown x-value is 0.10.



The calculator returns a value close to 260 days. Thus, 10% of all human pregnancies last fewer than about 260 days, and 90% last longer than 260 days. The value $x = 260$ is the cutoff separating the shortest 10% from the longest 90%. As introduced earlier, this cutoff is called the 10th percentile.

Example 3.6.4

Scores on a standardized reading test for 4th graders are normally distributed with mean $\mu = 514$ and standard deviation $\sigma = 117$.

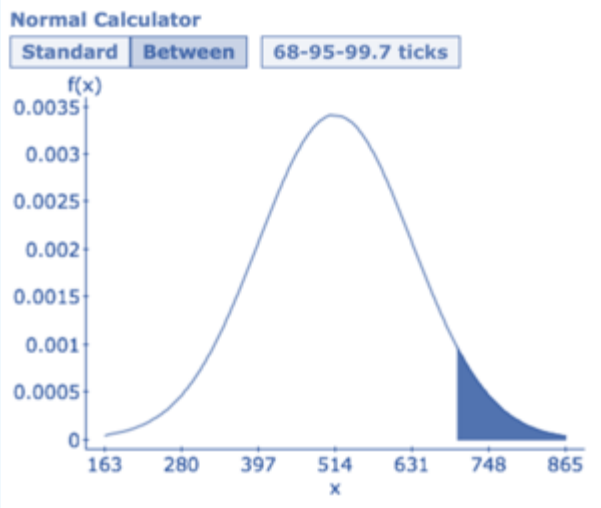
a. Find the percentage of students who score at least 700 on this test.

Solution

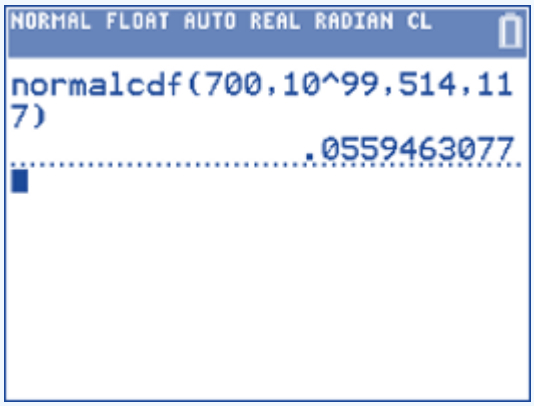
Translate the question into a mathematical statement:

$P(X > 700)$.

Sketch a normal curve centered at 514 with standard deviation 117. We shade the portion of the curve we are calculating: to the right of 700.



Use the **normalcdf(700, ∞, 514, 117)** command. The lower endpoint of the interval is 700 and the upper endpoint of the interval is +∞.



$P(X > 700) \approx 0.0559$. About 5.59% of 4th graders score at least 700 on this test.

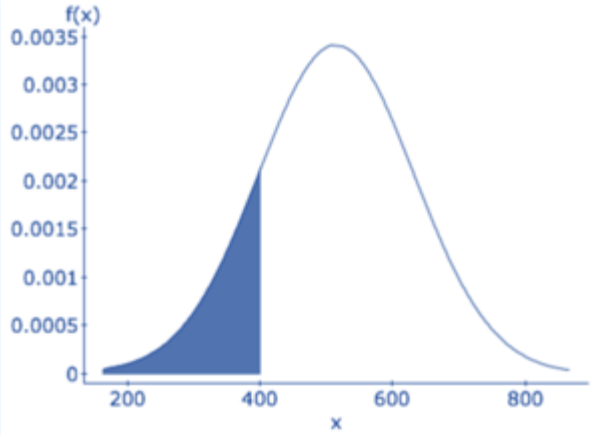
b. What percentage of 4th graders score less than 400 on the reading test?

Solution

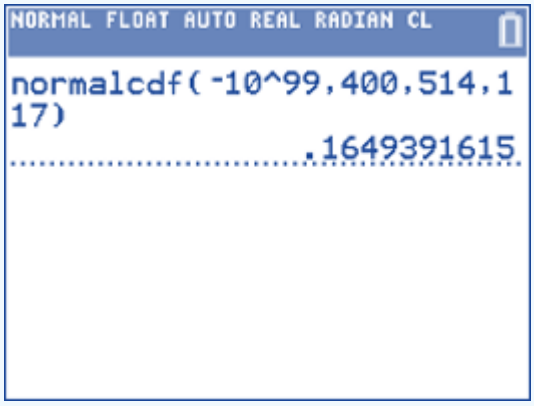
Translate the question into a mathematical statement:

$P(X < 400)$.

Sketch a normal curve centered at 514 with standard deviation 117. We shade the portion of the curve we are calculating: to the left of 400.



Use the **normalcdf(-∞, 400, 514, 117)** command. The lower endpoint of the interval is -∞ and the upper endpoint of the interval is 400.



$P(X < 400) \approx 0.1649$. Approximately 16.49% of 4th graders score less than 400.

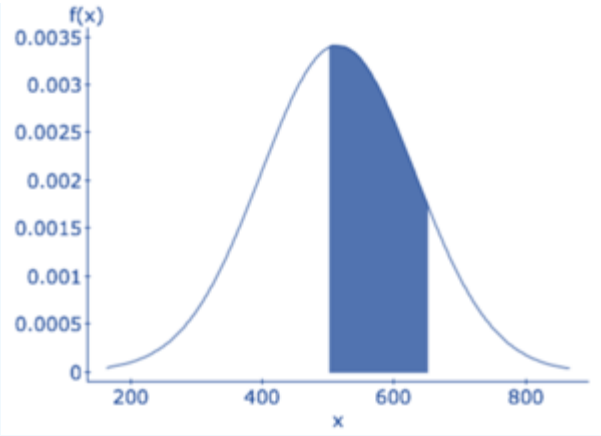
c. Find the percentage of 4th graders who score between 500 and 650, inclusive.

Solution

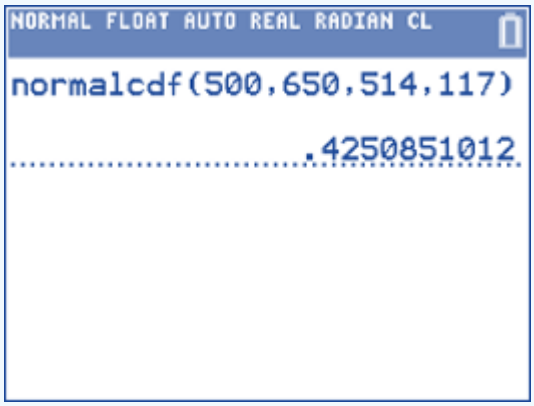
Translate the question into a mathematical statement:

$P(500 \leq X \leq 650)$.

Sketch a normal curve centered at 514 with standard deviation 117. We shade the portion of the curve we are calculating: between 500 and 650.



Use the **normalcdf(500, 650, 514, 117)** command. The lower endpoint of the interval is 500 and the upper endpoint of the interval is 650.



$P(500 \leq X \leq 650 \approx 0.4251$. About 42.51% of students score between 500 and 650.\

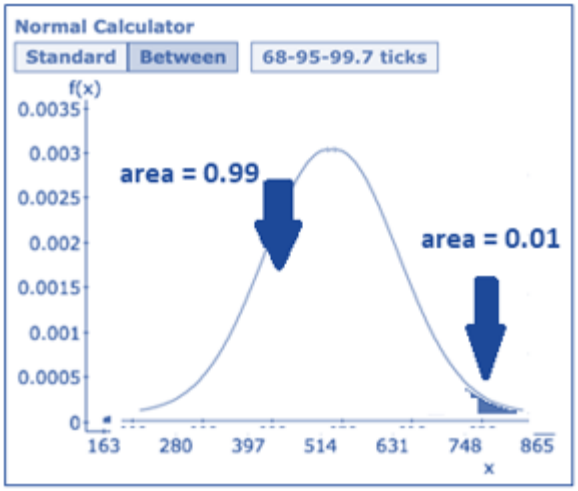
d. Students scoring in the top 1% qualify for a gifted program. What is the minimum score needed to qualify?

Solution

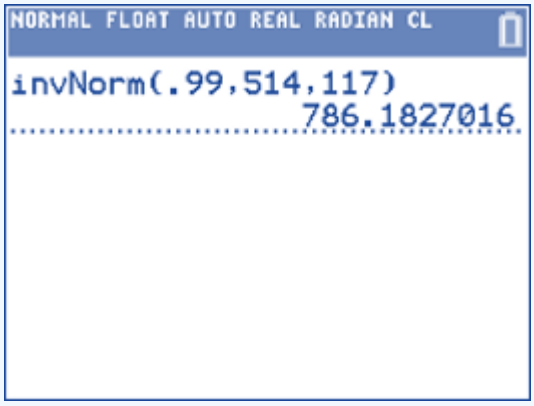
Translate the question into a mathematical statement:

$P(X \geq \ ?) = 0.01$.

Sketch a normal curve centered at 514 with standard deviation 117. We shade the right portion so that it contains 1% of the curve.



Use the **InvNorm(0.01, 514, 117)** command. If the area to the right of the cutoff value x is 0.01, then the area to its left is 0.99.



The result is $x \approx 786.18$. A student must score more than 786 points to place in the top 1%. This score corresponds to the 99th percentile.

You Try It $3.6.B$

The battery life of a certain brand of wireless headphones is approximately normally distributed with mean $\mu = 18.2$ hours and standard deviation $\sigma = 2.1$ hours.

Find the battery life (to the nearest tenth of an hour) such that 30% of all headphones last less than that amount.

Answer available at the end of this section.

You Try It $3.6.C$

At a two-year college, the number of hours students spend studying per week is approximately normally distributed with mean $\mu = 14.8$ hours and standard deviation $\sigma = 4.2$ hours.

Find the minimum number of study hours needed (to the nearest tenth of an hour) to be in the top 10% of all students.

Answer available at the end of this section.

Working with Quartiles and Percentiles

Recall from Section 2.2 that quartiles divide a distribution into four equal parts. Quartile 1 (Q_1) is the value such that 25% of the data lie below it and 75% lie above it. Quartile 3 (Q_3) is the value with 75% of the data below and 25% above.

Percentiles work in the same way but use 100 equal divisions instead of four. For example, if a score is at the 35th percentile, then 35% of all observations are less than that score and 65% are greater. Percentiles are widely used in standardized testing because they allow meaningful comparisons among students. If you need to review these concepts, please see Section 2.2.

Example 3.6.5

Scores on an IQ test are normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$.

a. Find the IQ scores that correspond to Q_1 and Q_3 Then, find the interquartile range.

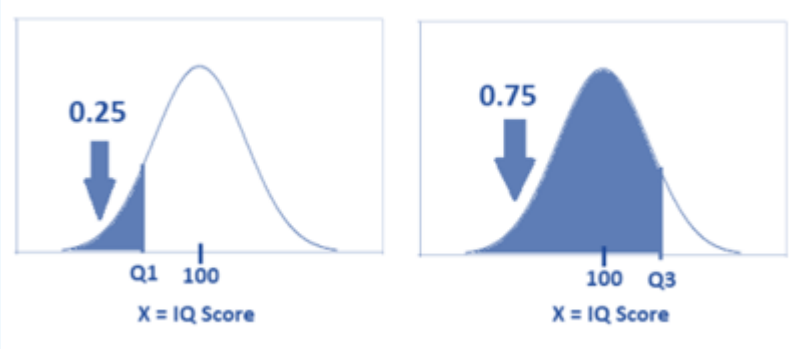
Solution

Q_1 is the value with 25% of the distribution to its left; Q_3 is the value with 75% to its left.

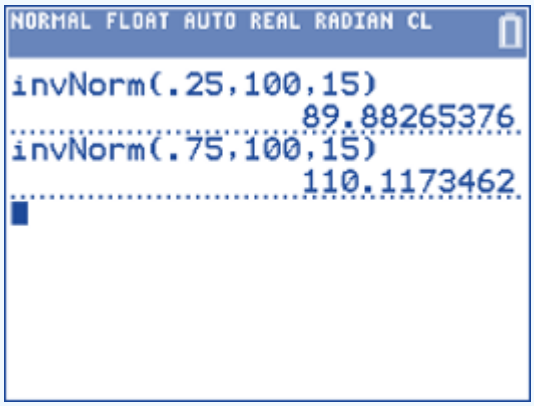
Translate into mathematical statements:

$P(Q_1 < \ ?) = 0.25$ and $P(Q_3 < \ ?) = 0.75$.

Sketch two normal curves centered at 100 with standard deviation 15. Shade one so that the portion to the left is 25%. Shade the other so that the portion to the left is 75%.



Use **InvNorm(0.25, 100, 15)** to find Q_1 and **InvNorm(0.75, 100, 15)** to find Q_3.



The quartiles are $Q_1 \approx 90$ and $Q_3 \approx 110$. The IQR is $Q_3 - Q_1 = 110 - 90 = 20$ points.

b. Jackson's IQ score was at the 65th percentile. What is Jackson's IQ score?

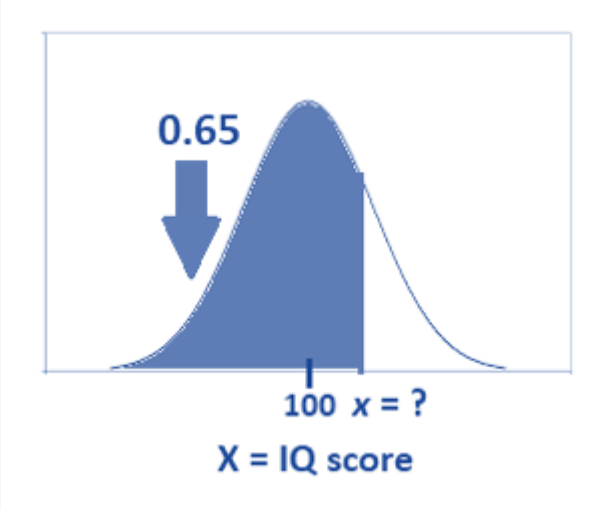
Solution

If Jackson is at the 65th percentile, then 65% of all scores fall at or below his score.

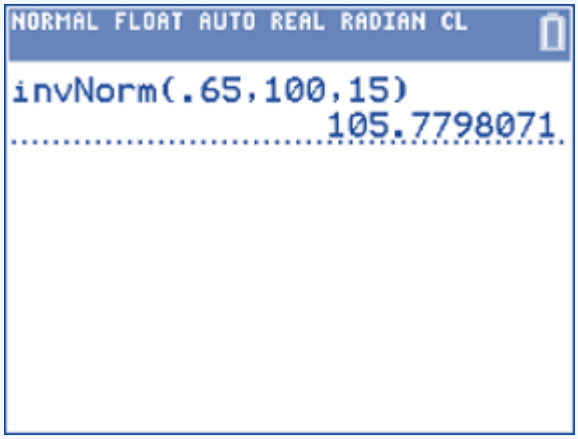
Translate the question into a mathematical statement:

$P(X \leq \ ?) = 0.65$.

Sketch a normal curve centered at 100 with standard deviation 15. Shade the left portion so that it contains 65% of the curve.



Use the **InvNorm(0.65, 100, 15)** command. The area to the left of the unknown x-value is 0.65.



The result is $x \approx 106$. Jackson's IQ score is approximately 106.

c. MENSA is an organization for high IQ members and requires an IQ in the top 2% of the population. Find the minimum IQ needed to qualify for the MENSA.

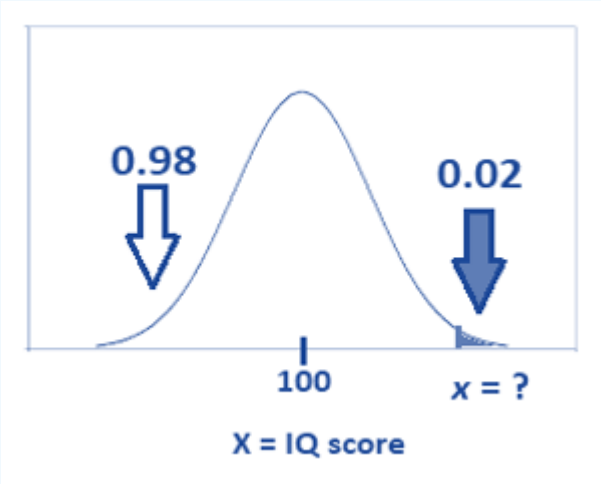
Solution

This means 2% of scores lie above the cutoff and 98% lie below it.

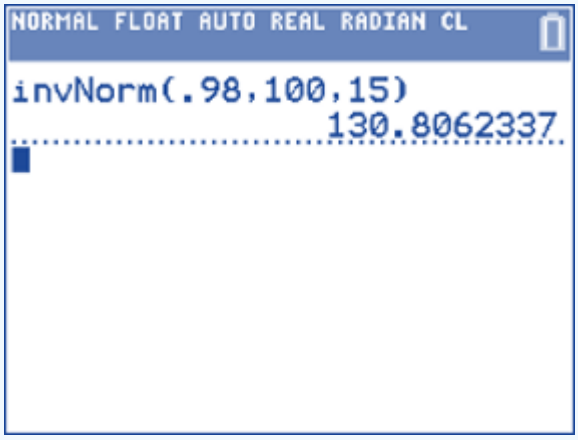
Translate the question into a mathematical statement:

$P(X \geq \ ?) = 0.02$.

Sketch a normal curve centered at 100 with standard deviation 15. Shade the right portion so that it contains 2% of the curve.



Use the **InvNorm(0.98, 100, 15)** command. If the area to the right of the cutoff value x is 0.02, then the area to its left is 0.98.



The result is $x \approx 131$. A person must have an IQ of at least 131 to qualify for MENSA. This corresponds to the 98th percentile.

You Try It $3.6.D$

The amount of coffee students drink per day is approximately normally distributed with mean $\mu = 14.2$ ounces and standard deviation $\sigma = 5.1$ ounces.

Find the amount of coffee (to the nearest tenth of an ounce) that corresponds to the 30th percentile of daily consumption.

Answer available at the end of this section.

You Try It $3.6.E$

The amount of time adults spend streaming video per day is $N(2.3,\ 0.9)$ hours.

Find the amount of streaming time (to the nearest tenth of an hour) that corresponds to Q_3.

Answer available at the end of this section.

With these tools, **normalcdf** for finding percentages and **invNorm** for finding cutoff values, you can solve a wide range of problems involving the normal distribution. The next example brings these ideas together in a single application.

Example $3.6.6$

Sarah keeps statistics for the baseball team. According to her records, the distances that fly balls travel are $N(250, 50)$ feet.

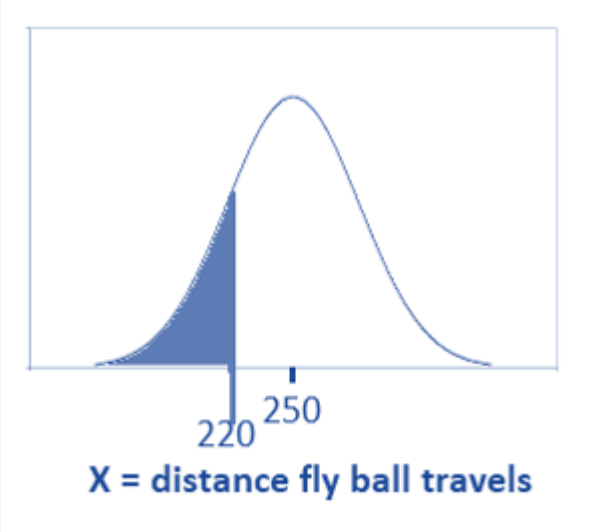
a. Based on Sarah's model, what percentage of fly balls travel less than 220 feet?

Solution

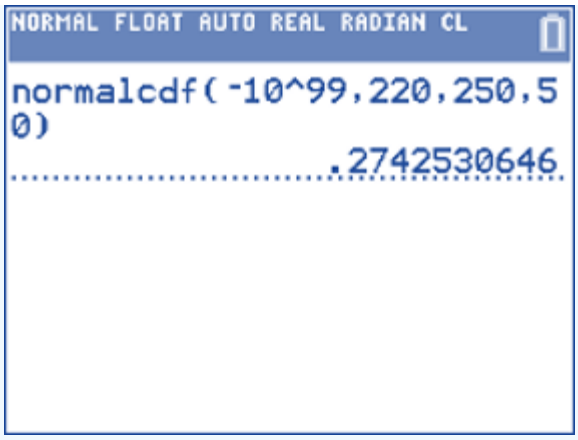
Translate the question into a mathematical statement:

$P(X < 220)$.

Sketch a normal curve centered at 250 with standard deviation 50. Shade the portion of the curve being calculated: to the left of 220.



Use the **normalcdf(-∞, 220, 250, 50)** command. The lower endpoint of the interval is -∞ and the upper endpoint of the interval is 220.



$P(X < 220) \approx 0.2743$. Approximately 27.43% of fly balls travel less than 220 feet.

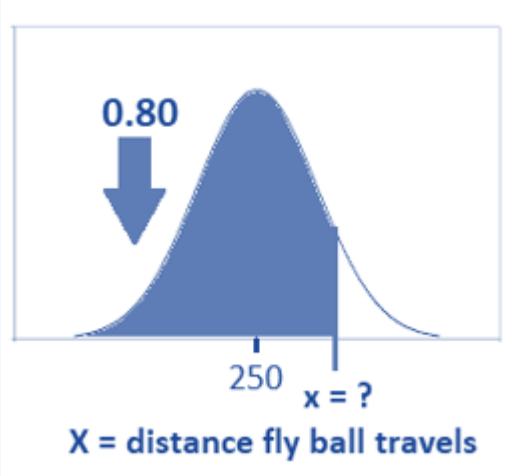
b. Find the 80th percentile.

Solution

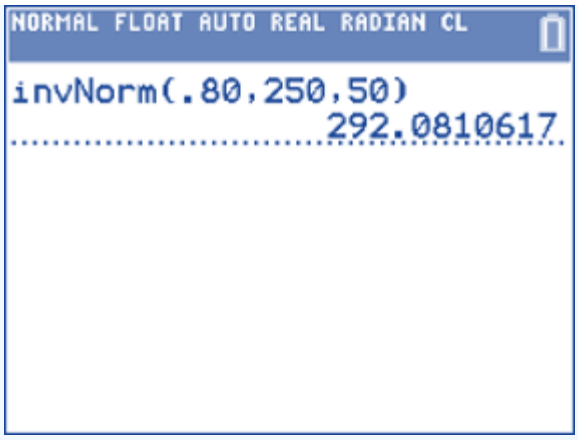
Translate the question into a mathematical statement:

$P(X \leq \ ?) = 0.80$.

Sketch a normal curve centered at 250 with standard deviation 50. Shade the left portion so that it contains 80% of the curve.



Use the **InvNorm(0.80, 250, 50)** command. The area to the left of the unknown x-value is 0.80.



The result is $x \approx 292$. The 80th percentile is about 292 feet. Roughly 80% of fly balls travel this distance or less.

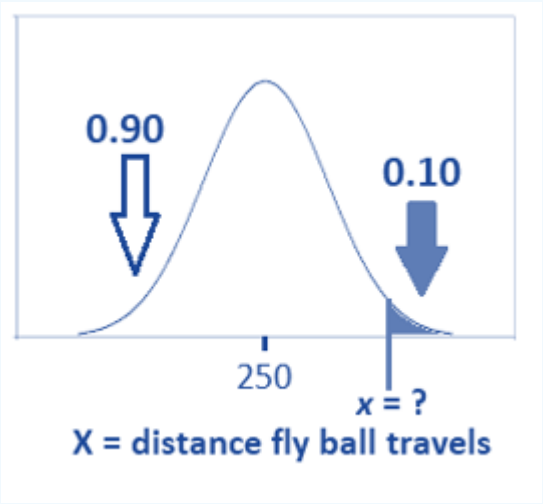
c. A player hit a fly ball so far that the distance it traveled ranked in the top 10% of all time. How far could the player have hit the fly ball?

This means 10% of scores lie above the cutoff and 90% lie below it.

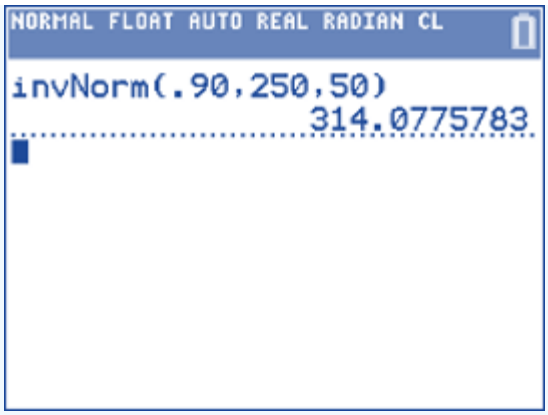
Translate the question into a mathematical statement:

$P(X \geq \ ?) = 0.10$.

Sketch a normal curve centered at 250 with standard deviation 50. Shade the right portion so that it contains 10% of the curve.



Use the **InvNorm(0.90, 250, 50)** command. If the area to the right of the cutoff value x is 0.10, then the area to its left is 0.90.



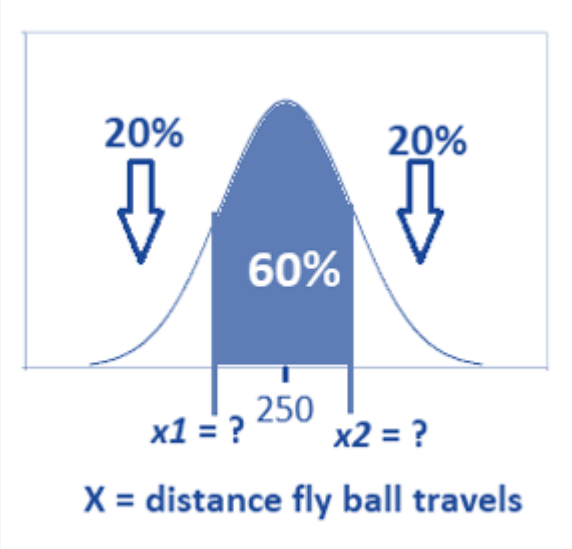
The result is $x \approx 314$. A fly ball must travel about 314 feet or farther to rank in the longest 10% of all time. This corresponds to the 90th percentile.

d. Give the range of distances that correspond to the middle 60% that all fly balls travel.

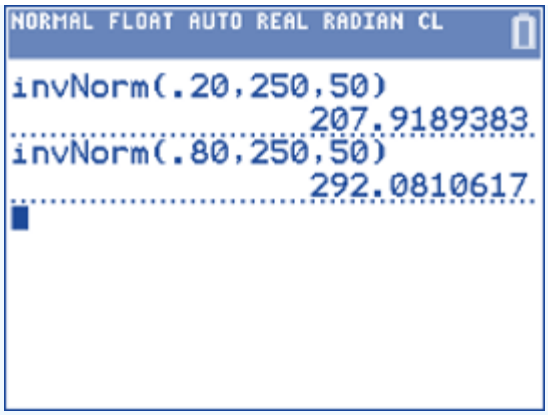
Translate the question into a mathematical statement:

$P(? \leq X \leq \ ?) = 0.60.$

Sketch a normal curve centered at 250 with standard deviation 50. Shade the middle portion so that it contains 60% of the curve. The middle 60% of the distribution lies between two cutoff values. The lower cutoff has 20% of distances to its left, and the upper cutoff has 80% to its left.



Use **InvNorm(0.20, 250, 50)** to find the first cutoff and **InvNorm(0.80, 250, 50)** to find the second cutoff.



The middle 60% of all fly balls travel between 208 feet and 292 feet.

You Try It $3.6.F$

The amount of time adults spend reading each day is approximately normally distributed with $\mu = 42$ minutes and $\sigma = 18$ minutes. Let

To the nearest whole minute, find the range of daily reading times that contains the middle 70% of all adults.

Answer available at the end of this section.

Section Summary

When solving problems involving a normal distribution, begin by translating the question into a clear mathematical statement and sketching the normal curve with the relevant region shaded. This helps you visualize the percentage or cutoff value you are trying to find.

To determine the percentage below, above, or between x-values, use normalcdf(on the TI-83/84 calculator.

To determine a cutoff x-value corresponding to a given left-side percentage, quartile, or percentile, use invNorm(. If the problem gives a right-sided percentage—for example, "top 15%"—you may need to convert it to a left-side percentage first by subtracting from 100%. Many older versions of the TI calculators require this left-side probability as the input to invNorm(.

Finally, these methods apply only when the variable of interest is approximately normally distributed. Having a reasonable justification for assuming normality is essential before using these tools. These tools will not work for distributions that are skewed, uniform, or otherwise not normal.

Section 3.6 Review Questions

1. The heights of women aged 18 to 24 are approximately normally distributed with mean 64.5 inches and standard deviation 2.5 inches.
   a. What percent of women in this age group are taller than 62 inches?
   b. What percent of women in this age group are between 65 and 70 inches tall?
   c. What percent of women in this age group are shorter than 59.5 inches?

2. Suppose yearly rainfall totals for a city in upstate New York follow a normal distribution, with mean 20 inches and standard deviation 5 inches. Find the percentage of years that total rainfall will be in each of the following intervals.
   a. less than 12 inches

b. greater than 25 inches
c. between 14 and 24 inches
d. greater than 40 inches

3. A study found that the length of Atlantic cod caught in nets have a mean 49.9 cm and a standard deviation 3.74 cm. Assume the length of fish is normally distributed.
   a. Find the percentage of Atlantic cod that are less than 52 cm long.
   b. Find the percentage of Atlantic cod that are more than 46 cm long.
   c. Find the percentage of Atlantic cod that are between 40.5 and 57.5 cm long.
   d. Determine the length such that 15% of all Atlantic cod are shorter than that value.

4. In the United States, males between the ages of 40 and 49 eat on average 103.1 grams of fat every day with a standard deviation of 4.3 grams. Assume the daily amount of fat that males eat is normally distributed.
   a. Draw a well-labeled sketch of the normal model.
   b. Use technology to find the percentage of all males aged 40-49 in the U.S. who eat
      i. more than 105 g of fat every day.
      ii. less than 95 g of fat every day.
      iii. less than 110 g of fat every day.
      iv. between 100 and 108 grams fat every day.
   c. Determine the amount of fat such that 15% of all males aged 40–49 eat more than that amount.

5. A citrus farmer who grows tangerines finds that the diameters of tangerines harvested on his farm follow a normal distribution with a mean diameter 5.85 cm and standard deviation of 0.24 cm.
   a. Find the percentage of tangerines from this farm with a diameter larger than 6.0 cm.
   b. The middle 20% of tangerines from this farm have diameters between __ and __ cm.
   c. Find the 90th percentile for the diameters of tangerines and interpret it in a complete sentence.

6. Scores on the ACT for a certain group of test takers have a mean of approximately 18 and standard deviation of 6. These scores follow a normal distribution.
   a. What percentage of ACT math scores are higher than 20?
   b. John takes the ACT math test. He scores 26. At what percentile did John score?
   c. Susan doesn't remember her score, but she knows she scored at the 60th percentile. What was her score?

7. The wrapper of a certain candy bar lists its weight as 2.13 ounces. Naturally, the weights of individual bars vary somewhat. Suppose that the weights of these candy bars vary according to a normal distribution, with $mu = 2.2$ ounces and $\sigma = 0.04$ ounces.
   a. What proportion of the candy bars weigh less than the advertised weight?
   b. What proportion of the candy bars weigh between 2.2 and 2.3 ounces?
   c. About 90% of these candy bars weigh less than what amount?

8. In the US, the age of smartphone users follows an approximate normal distribution with mean and standard deviation of 36.9 years and 13.9 years, respectively.
   a. What percentage of smartphone users are younger than 25 years old?
   b. Calculate the interquartile range (IQR) for the ages of smartphone users.
   c. Forty percent (40%) of smartphone users are at least how old?

Answers

You Try It $3.6.A$
a. 31%
b. 76%

You Try It $3.6.B$
16.4 hours

You Try It $3.6.C$
20.2 hours

You Try It $3.6.D$
11.0 ounces

You Try It $3.6.E$

2.9 hours

You Try It $3.6.F$

23 to 61 minutes

This page titled 3.6: Applying a Normal Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

4: CORRELATION AND REGRESSION

4.1: Scatterplots and Correlation
4.2: Interpreting a Least Squares Regression Model
4.3: Building a Least Squares Model

This page titled 4: Correlation and Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

4.1: Scatterplots and Correlation

Learning Objectives

- Use scatterplots to display bivariate data.
- Describe correlation as shown in scatterplots.
- Interpret a correlation coefficient.
- Calculate the correlation coefficient.
- Apply properties and cautions about correlation.

Introduction

Up to this point, we have focused on exploring single variables -- how to describe their distributions numerically, how to visualize their shapes, and how to interpret what those patterns tell us. This gives us a strong foundation for analyzing data, but real-world questions rarely involve just one variable in isolation. In everyday situations, variables often move together in meaningful ways.

Think about common examples: as the number of hours a student works increases, their weekly free time might decrease; as outdoor temperature rises, ice-cream sales tend to rise as well; as practice time increases, performance often improves. When two quantitative variables appear connected, we naturally want to understand how they are related and how strongly they move together.

This next section introduces the idea of studying pairs of quantitative variables. Our goal is to describe the form, direction, and strength of the relationship between them. When the connection is strong enough, knowing one variable can help us make informed predictions about the other. To quantify this relationship, we use the concept of correlation -- a numerical measure that captures how two quantitative variables change together.

As we move forward, we will learn how to visualize relationships between variables, how to compute correlation, and how to interpret what correlation does (and does *not*) tell us about the underlying connection.

Bivariate Data and Scatterplots

When two quantitative variables are measured on the same individuals, we call the resulting information **bivariate data**. Each subject contributes a *pair* of observations, one for each variable. For example, consider the data set below, which records verbal SAT scores and college GPA for seven students. Each student has two associated measurements, forming a pair of values.

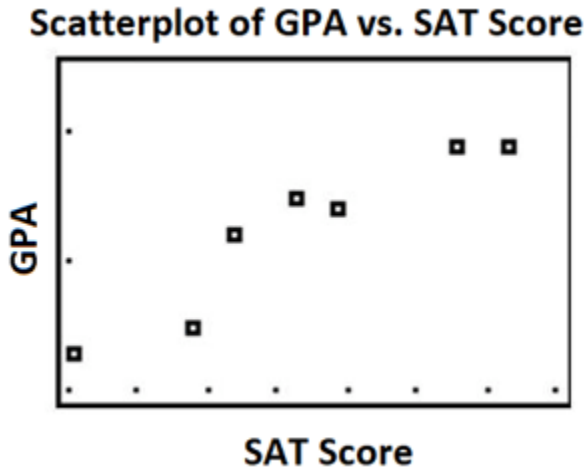
Verbal SAT Score and GPA for 7 Students



When we examine these paired observations carefully, a pattern begins to emerge: students with higher SAT scores tend to have higher GPAs, and students with lower SAT scores tend to have lower GPAs. This suggests a relationship between the two variables.

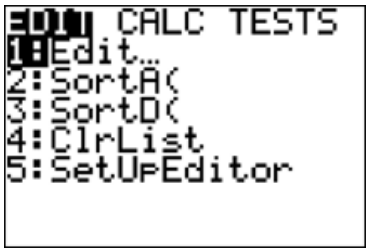
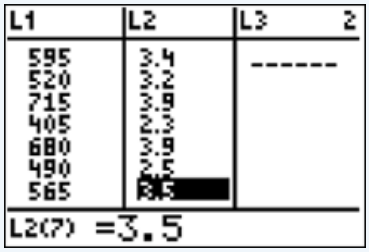
To look for relationships visually, we can use a **scatterplot**, one of the most important tools for analyzing bivariate data. In a scatterplot, each point represents a single subject, plotted using its pair of values, one on the horizontal axis and one on the

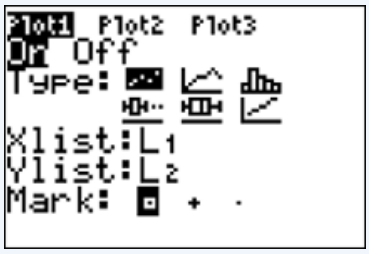
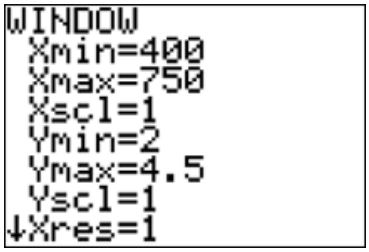
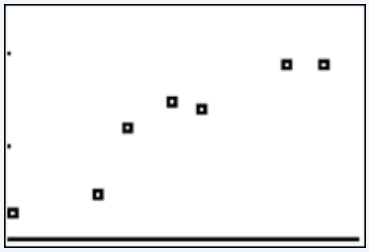
vertical axis. Scatterplots allow us to see trends, clusters, unusual observations, and whether a linear pattern is present.



Although scatterplots can be drawn by hand, technology makes the process faster and more precise. The TI-84 calculator, for example, can generate scatterplots efficiently. The next section provides step-by-step instructions for creating a scatterplot on the TI-84 and preparing the data for further analysis.

Drawing a Scatterplot on the TI-83/84 Calculator

1.	Press **[STAT].** On the **EDIT** menu, **1:Edit** and press **[ENTER]**.	
2.	Enter the data, with the SAT Scores in L1 and the GPAs in L2.	

3.	Press **[2nd] [Y=]** to enter the **STAT PLOTS** menu, and choose **1:Plot 1.**	
4.	Change the Plot settings to match the following: Make sure the graph is set to **On**. For **Type**, choose the scatterplot icon. For **Xlist**, choose L1 and for **Ylist** choose L2. This selects a scatterplot with SAT scores along the x-axis and the GPA along the y-axis. To see the points better, you should choose either the square or the plus sign for the mark. The square has been chosen in the screenshot.	
5.	Set the **[WINDOW]** to match the data. In this case, we look at our lowest and highest data values in each variable and added a bit of room to create a pleasant window.	
6.	Press **[GRAPH]** to see the scatterplot.	
	Note: Rather than manually setting the window in Step 5, you can let the calculator automatically set the window based on your data. Press **[ZOOM]** and select **9:ZoomStat**. The values used for the window will be slightly different that the ones we chose manually, but the graph will look essentially the same.	

Patterns in Scatterplot Graphs

When examining bivariate data, it is essential to identify several key features of any relationship that may exist between the two variables: **form**, **direction**, **strength**, and the presence of **unusual observations**. These features help us understand how the variables behave together and prepare us for a more formal discussion of *correlation* later in this section.

Before we define correlation precisely, it is helpful to think about it informally. Correlation gives us a sense of how two quantitative variables tend to vary together. If knowing the value of one variable helps us anticipate the value of the other -- even loosely -- then the variables exhibit some degree of correlation. At this stage, we are simply developing intuition: correlation reflects the overall pattern in how paired data points move together, without yet assigning a numerical value.

Form: Linear vs. Nonlinear Relationships

The first question to ask when viewing a scatterplot is: *What is the form of the relationship?* Understanding the form tells us how we can proceed in quantifying the relationship and whether tools such as correlation and linear regression are appropriate.

A **linear relationship** occurs when the points cluster around a straight line—either sloping upward or downward. Linear patterns are the focus of correlation because correlation specifically measures the strength and direction of a linear association.

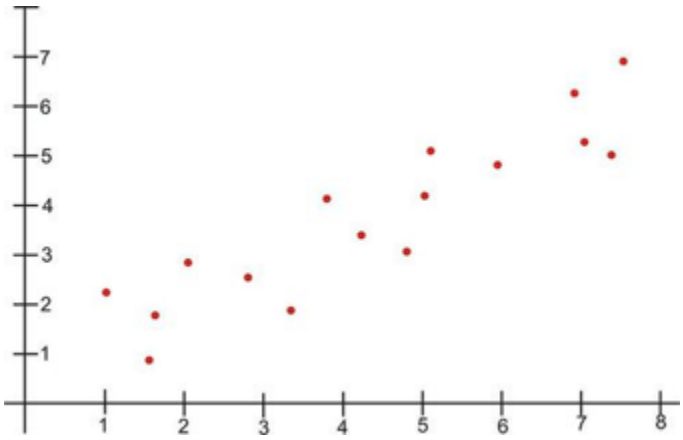
However, not all relationships are linear. Some scatterplots show **nonlinear patterns**, such as curves, U-shapes, or exponential growth. In these cases, even if the variables are strongly related, correlation may not capture the relationship accurately because it only describes linear trends.

Direction: Positive, Negative, or Zero

Once we determine that the form, we examine the direction of the relationship.

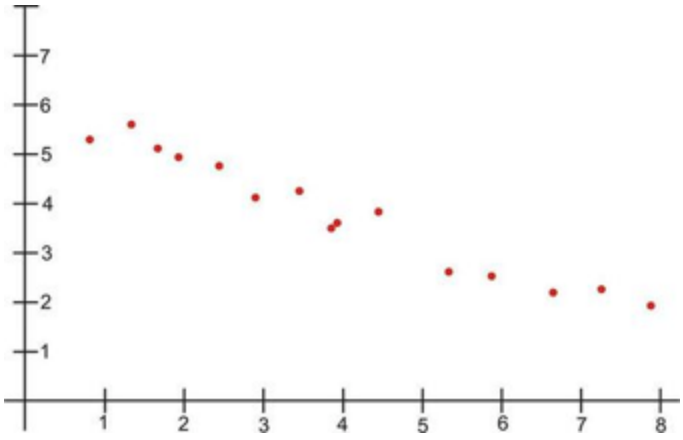
Positive Relationship

When the points in a scatterplot tend to rise from the lower left toward the upper right, the variables show a **positive relationship**. In this pattern, higher values of one variable are generally associated with higher values of the other, and lower values are associated with lower values. Many real-world relationships—such as hours studied and exam scores, or height and weight—often display this type of trend.



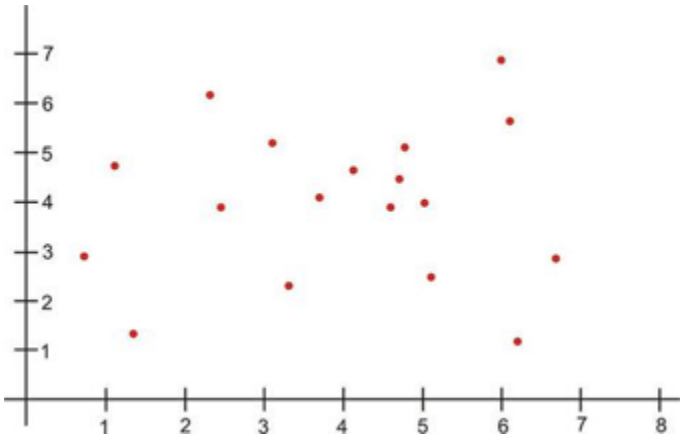
Negative Relationship

When the points tend to fall from the upper left toward the lower right, the variables show a **negative relationship**. In this case, higher values of one variable are associated with lower values of the other. Examples include the relationship between the number of absences and course grade, or speed and travel time.



No Relationship

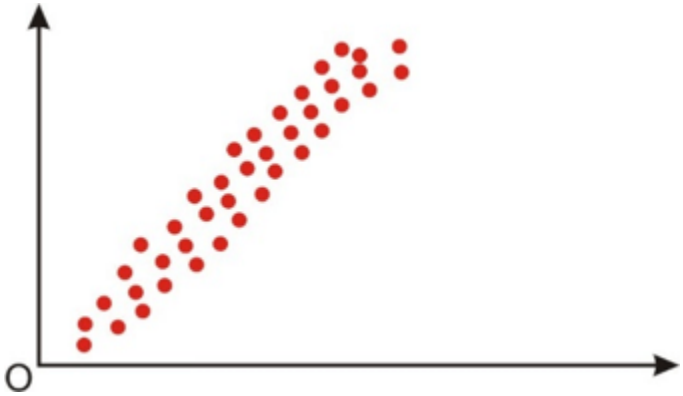
If the points show no clear upward or downward trend, the variables exhibit **no relationship** (or nearly none). The points may form a cloud with no discernible direction. In such cases, knowing one variable does not help predict the other.



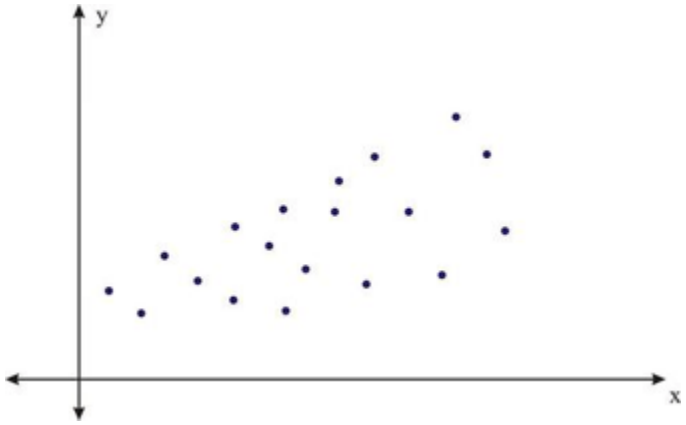
Strength of a Relationship

Describing a relationship is not only about direction—it also involves **strength**. Strength describes how tightly the points cluster around the form you identified.

- If the points are **close together**, forming a narrow, elongated cluster that follows a discernable form, the relationship is strong. Here is an example of a positive linear relationship that is quite strong:



- If the points are **widely scattered**, forming a broad, loose cloud, the relationship is weak. Here is an example of a another positive linear relationship, but it is much weaker than the one above:



Unusual Observations: Outliers and Influential Points

Scatterplots also help us identify unusual observations. An **outlier** is a point that does not follow the general pattern of the data. Outliers may occur because of measurement error, unusual circumstances, or natural variability. A single outlier can weaken or distort the apparent relationship.

Example 4.1.1

For each pair of variables, decide whether the relationhip would likely be positive, negative, or close to 0. Base your reasoning on how the variables tend to vary together in real-world contexts.

a. hours of sunlight and solar panel energy output

Solution

Positive: More sunlight provides more energy for solar panels to convert into electricity. As sunlight increases, energy output tends to increase in a predictable way.

b. number of cups of coffee consumed and hours of sleep obtained

Solution

Negative: Drinking more coffee—especially later in the day—tends to reduce the amount of sleep a person gets. As coffee consumption increases, sleep hours often decrease.

c. distance driven per day and amount of gasoline remaining in the tank

Solution

Negative: The farther someone drives, the more fuel is used. As daily driving distance increases, the amount of gasoline left in the tank tends to decrease.

d. student's height and score on a history exam

Solution

Close to zero: There is no meaningful reason to expect taller students to perform better or worse on a history exam. The variables do not vary together in any predictable way.

Correlation Coefficient

Once we have examined a scatterplot and described the form, direction, strength, and any unusual observations, we can summarize a linear relationship between two quantitative variables more precisely using a numerical measure called the **correlation coefficient**. The correlation coefficient, symbolized by r, is a statistic that quantifies the strength and direction of a **linear** relationship between two variables.

Correlation is always a number between –1.0 and +1.0:

- A positive value for r indicates a positive linear relationship: as one variable increases, the other tends to increase.
- A negative value for r indicates a negative linear relationship: as one variable increases, the other tends to decrease.
- A value of 0 for r indicates no *linear* relationship, although a nonlinear relationship may still exist.

The absolute value of r, written $|r|$, describes the strength of the linear relationship. Values closer to 1 indicate a stronger linear association; values closer to 0 indicate a weaker one. Although interpretations vary somewhat across disciplines, the following guidelines are commonly used:

- If $|r|$ is between 0.80 and 1.0, the correlation is strong.
- If $|r|$ is between 0.50 and 0.80, the correlation is moderate.
- If $|r|$ is between 0.10 and 0.50, the correlation is weak.
- If $|r|$ is less than 0.10, there is no apparent linear correlation.

A correlation of +1.0 represents a perfect positive linear relationship, and a correlation of –1.0 represents a perfect negative linear relationship. These occur only when all points lie exactly on a straight line. Data collected from a sample will rarely, if ever, be so ideal.

It is important to emphasize that correlation measures only linear relationships. A correlation coefficient of 0 does not necessarily mean the variables are unrelated; they may follow a curved pattern such as a quadratic or exponential relationship. This is why examining the scatterplot first is essential.

How the Correlation Coefficient Is Computed

Although technology now computes r instantly, understanding how it is calculated helps clarify what the statistic represents. Correlation is based on z-scores, which measure how far each observation is from its mean in standard deviation units.

To compute r for variables x and y:

1. Convert each x value to a z-score, z_x.
2. Convert each y value to a z-score, z_y.
3. Multiply the paired z-scores for each individual: $z_x \cdot z_y$.
4. Add all these cross products together.
5. Divide the sum by $n - 1$, where n is the number of individuals.

This produces the formula for the correlation of variable x with variable y:

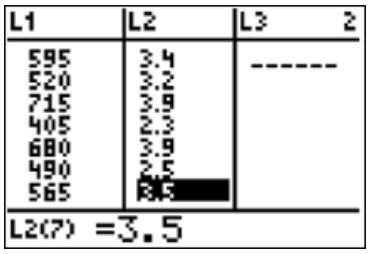
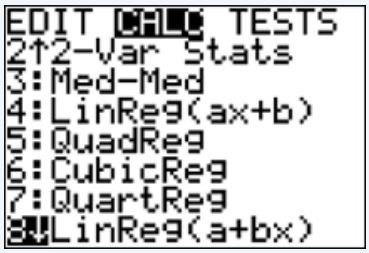
$$r_{xy} = \frac{\dsum z_x z_y}{n-1}.$$

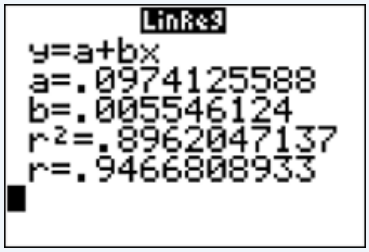
Conceptually, r is the average product of paired z-scores. When high z-scores for x tend to pair with high z-scores for y (or low with low), the products are positive, producing a positive correlation. When high z-scores for x pair with low z-scores for y, the products are negative, producing a negative correlation.

Using Technology to Compute r

In practice, correlation is almost never computed by hand! Tools such as the TI-83/84 calculator, spreadsheets, and statistical software can compute r quickly and accurately. The table below provides instructions for using the TI-84 to find the correlation coefficient for the SAT Score and GPA data introduced earlier.

Finding the Correlation Coefficient on the TI-83/84 Calculator

1.	Press **[STAT]** and choose **1:Edit**. Input the data, with the SAT Scores in L1 and the GPA data in L2 as you did for the scatterplot.	
2.	Press **[STAT]** again. Use the arrow to scroll over to the **CALC** menu. Choose *8:LinReg(a+bx)*.	
3.	You will have one of the two screens shown here. If you have the top one, the blinking cursor is waiting for you to enter the names of the two lists of data separated by a comma: L1,L2. Then, press **[ENTER]**. If you have the STAT WIZARD screen like the one on the bottom, complete the entries for **Xlist** and **Ylist** with L1 and L2. Leave **FreqList** and **Store RegEQ** blank for now. Highlight **Calculate** and press **[ENTER]**.	 

4.	The correlation coefficient shows up as *r*. These other values will be discussed in a later section.	
	Note: If the value r^2 and *r* are missing, make sure that your **Diagnostics** are turned on in the menu under **[MODE]**.	

The correlation coefficient is $r \approx 0.95$. This value indicates a very strong linear association between SAT score and GPA. The scatterplot supports this conclusion as well, showing that the paired data points closely follow a linear pattern that slopes upwards.

Example 4.1.2

Anthropologists sometimes estimate a person's height from skeletal remains. One useful measurement is the length of the first metacarpal bone. To study this relationship, an anthropologist collected paired data on *metacarpal length* (cm) and *height* (cm) for several individuals.

Using technology, create a scatterplot, describe any relationship you observe, and confirm your description using the correlation coefficient.

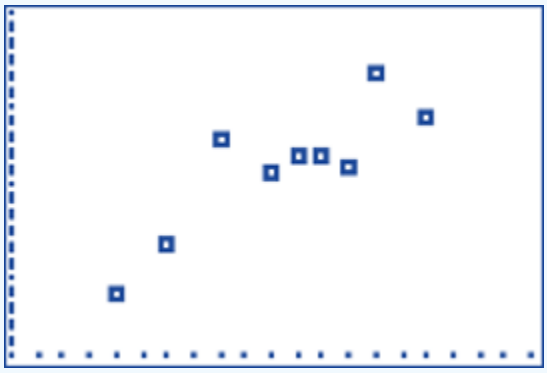
Metacarpal versus Height

Person	Length of Metacarpal (cm)	Height of Person (cm)
A	45	171
B	51	178
C	39	157
D	41	163
E	48	172
F	49	183

Person	Length of Metacarpal (cm)	Height of Person (cm)
G	46	173
H	43	175
I	47	173

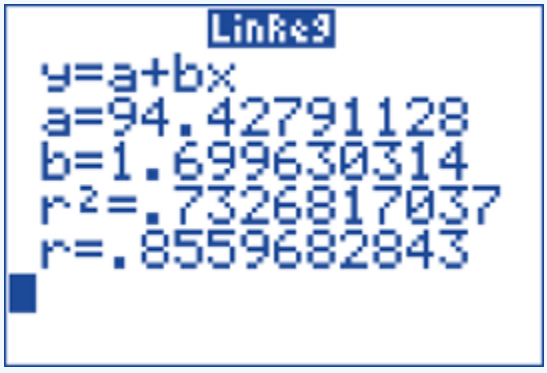
Solution

A scatterplot of the bivariate data, displayed on a window of [35,55] for metacarpal length and [150,190] for height, is shown below.



Based on the graph, the relationship between metacarpal length and height appears positive, strong, and linear. The points fall within a narrow, elongated cluster with relatively little scatter, indicating a clear linear trend. As metacarpal length increases, height tends to increase as well.

Because the scatterplot shows a tight linear pattern, we expect the correlation coefficient to be fairly large in magnitude. The calculator output confirms this:



The correlation coefficient is $r \approx 0.86$, which is a positive value with a magnitude close to 1. This indicates a strong positive linear association, consistent with what we observed in the scatterplot.

Example 4.1.3

Rearrange these correlation coefficients in order from weakest to strongest:

$$0.79, \quad -0.43, \quad -0.40, \quad -0.82, \quad 0.08$$

Solution

The strength of a linear relationship is determined by the absolute value of the correlation coefficient. The closer $|r|$ is to 1, the stronger the linear relationship; the closer $|r|$ is to 0, the weaker the relationship.

Ordering the coefficients from weakest to strongest based on their absolute values:

$$0.08, \quad -0.40, \quad -0.43, \quad 0.79, \quad -0.82.$$

The sign of the correlation indicates the direction of the relationship, but it does not affect the strength. Strength depends solely on the magnitude of the correlation, $|r|$.

Properties of Correlation

Now that we have calculated the correlation coefficient and discussed the meaning of its sign and absolute value, we can summarize several important properties of correlation. These properties help us understand how correlation behaves and how it should -- and should not -- be interpreted.

- **Non-resistant measure** — The correlation coefficient is highly sensitive to outliers. Even a single unusual observation can substantially increase or decrease the value of r. This is why examining the scatterplot for unusual points is essential before interpreting correlation.
- **Symmetry** — Correlation treats both variables equally. The correlation between x and y is the same as the correlation between y and x. Switching the horizontal and vertical axes in a scatterplot does not change the value of r.
- **Unitless measure** — Correlation has no units. Whether height is measured in centimeters or inches, or income in dollars or euros, the correlation coefficient remains the same. It is always reported as a number between –1 and 1, not as a percentage.
- **Scale invariance** — Changing the scale of either variable—such as converting temperatures from Celsius to Fahrenheit or rescaling test scores—does **not** affect the correlation. Adding a constant or multiplying by a positive constant leaves r unchanged. This property reflects the fact that correlation depends on standardized values (z-scores), not on the original units.

These properties highlight why correlation is a powerful descriptive statistic: it summarizes the direction and strength of a linear relationship in a way that is unaffected by units or scale. At the same time, its sensitivity to outliers reminds us that correlation must always be interpreted alongside a scatterplot.

Correlation ≠ Causation

It is often tempting to conclude that when two variables show a strong correlation, one variable must be causing the other to change. This is a mistake we must avoid. Correlation simply tells us that two variables tend to vary together in a predictable way. It does not tell us *why* they vary together.

A strong linear relationship may arise for many reasons that have nothing to do with cause and effect. In many situations, a third variable -- or even a combination of other factors -- may be influencing both variables at the same time. When this happens, the two correlated variables appear connected, but neither one is actually causing the other to increase or decrease. For this reason, it is essential to interpret correlated variables as related, not causal, unless additional evidence supports a cause-and-effect conclusion.

Example 4.1.4

There is a strong correlation between *shoe size* and *reading ability* in children. It would be unreasonable to think that larger feet somehow make a child read better, or that reading more causes feet to grow. Instead, both variables change in predictable ways because of a third factor: *age*. As children grow older, their feet get larger and their reading skills improve. Age explains the relationship, not a causal link between shoe size and reading ability.

This example illustrates why correlation alone cannot establish causation. To make causal claims, we need additional evidence—such as controlled experiments or well-designed observational studies—that goes beyond the correlation itself.

Section Summary

Bivariate data consist of paired observations recorded for the same subject. When we study bivariate data, we are interested in how the two variables vary together. Scatterplots allow us to visualize these paired observations and describe the form, direction, and strength of any linear relationship. From a scatterplot, we can see whether the relationship appears linear, whether the variables tend to move together or in opposite ways, and whether the points cluster tightly around a line or show substantial scatter. We also can observe whether one or more individuals fail to belong to the general trend as an outlier.

The correlation coefficient, r, provides a precise numerical measure of the direction and strength of a linear relationship. Correlation values range from –1.0 to +1.0, with values near ±1 indicating a strong linear association and values near 0 indicating little to no linear association. Correlation is unitless and scale-invariant. It is typically computed using technology.

Several factors can influence the value of the correlation coefficient. Outliers can distort correlation substantially because r is a non-resistant measure. Nnonlinear relationships may produce correlations near zero even when the variables are strongly related. For these reasons, correlation should always be interpreted alongside a scatterplot and with attention to the context of the data.

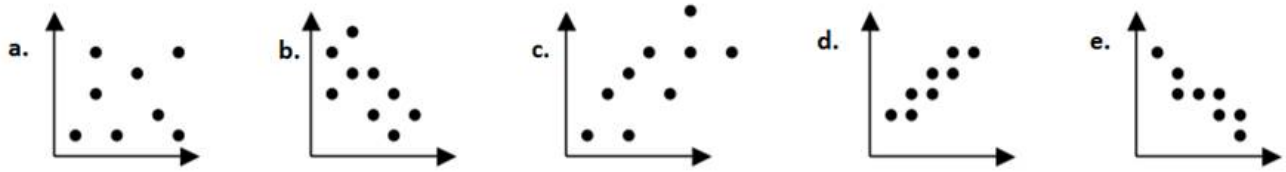
Finally, it is important to remember that correlation does not imply causation. A strong correlation indicates that two variables vary together in a predictable way, but it does not tell us why they vary together. A third variable—or a combination of factors—may be responsible for the observed relationship. Correlation describes association, not cause-and-effect.

Section 4.1 Review Questions

1. For each pair of variables tell whether you think their correlation would be positive, negative, or close to zero.
   a. temperature outside and heating cost
   b. length of hair and amount of shampoo used
   c. a runner's speed and the time needed to finish a race
   d. college GPA and shoe size
   e. amount of rain and number of umbrellas sold
   f. amount of rain and number of lawn sprinklers sold
   g. amount of rain and number of children absent from school

2. Sketch an example of a scatterplot with at least eight pairs of data in which the correlation is
   a. strong but not perfect, and negative
   b. weak and positive

3. Tell whether each scatterplot shows *positive correlation, a negative correlation*, or *no correlation*.



4. Using the same graphs as in question 3, identify the value of r from the list of correlation coefficients below that best matches each scatterplot:

   $r = -0.78$ $r = 0.92$ $r = -0.05$ $r = 1.0$

   $r = 1.53$ $r = 0.75$ $r = -1.0$ $r = -0.95$

5. For each correlation coefficient, describe what it means for data to have that correlation coefficient and roughly sketch a scatterplot with at least eight pairs of data having that correlation coefficient.
   a. $r = 1$
   b. $r = -0.8$
   c. $r = -1$
   d. $r = 0$
   e. $r = 0.5$

6. These observations were taken for five students measuring grade and reading level.

Grade vs. Reading Level for Five Students

Student Number	Grade	Reading Level
1	2	6
2	6	14
3	5	12
4	4	10
5	1	4

   a. Draw a scatterplot for these data.
   b. Use the scatterplot to describe the relationship between grade and reading level.
   c. Calculate the correlation coefficient.

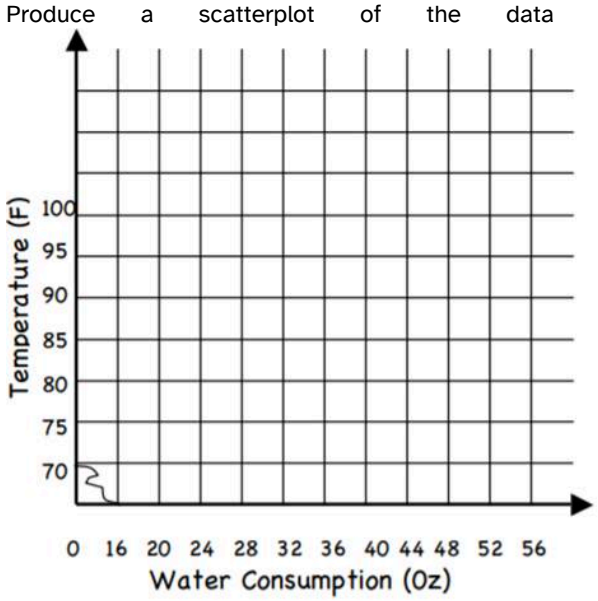
7. Nathaniel records his water consumption and the air temperature during a 3-hour period on 7 randomly selected days during the summer. Nathan's data are shown in the table.

Problem 7: Water Consumption and Air Temperature

Water Consumption (ounces)	Temperature (degrees F)
48	99

Water Consumption (ounces)	Temperature (degrees F)
27	85
48	80
16	75
32	92
25	85
20	83

a. Produce a scatterplot of the data on a graph similar to the one provided.



b. Analyze the relationships in terms of form, direction, strength, and outliers.
c. Produce the scatterplot on a graphing calculator with the same window setting as above.
d. Use the calculator to find the correlation coefficient.
e. Use the calculator to find the correlation coefficient again but this time exclude the outlier. Is correlation a resistant measure?

8. A teacher gives two quizzes to his class of 10 students. The following are the scores of the 10 students.

Problem 8: Quiz Results for Ten Students

Student	Quiz 1	Quiz 2
1	15	20
2	12	15
3	10	12
4	14	18

Student	Quiz 1	Quiz 2
5	10	10
6	8	13
7	6	12
8	15	10
9	16	18
10	13	15

a. Use technology to create a scatterplot.
b. In a sentence or two, describe the relationship between the scores on the two quizzes.
c. Compute the correlation coefficient *r* between the scores on the two quizzes.

This page titled 4.1: Scatterplots and Correlation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

4.2: Interpreting a Least Squares Regression Model

Learning Objectives

- Identify explanatory and response variables.
- Predict values using a regression line.
- Interpret slope and *y*-intercept of a regression line in the context.
- Calculate residuals in regression analysis.

Introduction

In Section 4.1, we used the correlation coefficient to describe the strength and direction of a linear relationship between two quantitative variables. Scatterplots helped us visualize these relationships and recognize patterns in how paired data values tend to vary together. While correlation provides a useful summary of how two variables are related, it does not allow us to make precise predictions. Correlation tells us that two variables vary together, but not how much one variable is expected to change when the other changes. To answer questions of that kind, we need a method that goes beyond describing a relationship and begins to use that relationship for estimation.

In this section, we turn to a technique designed specifically for prediction -- the process of estimating the value of one variable based on the value of another using a clear trend in the data. This approach will allow us to move from general descriptions of association to concrete, numerical predictions grounded in the observed pattern of the data.

Explanatory and Response Variables

Once we have identified two variables that are related, we can begin developing a prediction model for that relationship. In such a model, one variable plays the role of the **explanatory variable**, and the other becomes the **response variable**. The explanatory variable is used to help predict or explain changes in the response variable.

When graphing bivariate data for prediction, the explanatory variable is placed on the x-axis, and the response variable is placed on the y-axis. In other contexts, these are sometimes called the *independent* and *dependent* variables, respectively.

When we are simply analyzing bivariate data for correlation, we do not need to distinguish between explanatory and response variables. Correlation treats both variables symmetrically, so it does not matter which variable is labeled x or y. However, once we shift from describing a relationship to predicting one variable from another, the distinction becomes essential. In a prediction model,

- y represents the response variabl - the variable we want to predict, and
- x represents the explanatory variable -- the variable used to make the prediction.

How do we decide which variable should be x and which should be y? The decision depends on the context and the purpose of the analysis as we will see in the following examples.

Example 4.2.1

Someone tells you that the number of cricket chirps per minute can be used to estimate the outside temperature.

Here, the *temperature* is the response variable because it is the value we want to predict. The *number of chirps* is the explanatory variable because it is used to make that prediction.

You Try It 4.2.A

Jen plants 40 seedlings and exposes them to different amounts of sunlight (measured in hours per day). After a month, she measures each seedling's growth. She uses the amount of sunlight each plant gets to estimate how much it will grow.

What variable is most likely Jen using as the response variable and which is the explanatory variable?

Answer available at the end of this section.

When the Choice Is Not Obvious

Sometimes the relationship between two variables does not clearly suggest which one explains the other. In these cases, the choice depends on the goal of the analysis.

Consider a situation where Tim and Kelly collect data on their classmates' heights and weights.

- Tim wants to describe the relationship between *height* and *weight*. In this case, there is no obvious explanatory or response variable; Tim is simply studying the relationship.
- Kelly wants to predict a student's *weight* from their *height*. Here, *height* becomes the explanatory variable, and *weight* becomes the response variable.

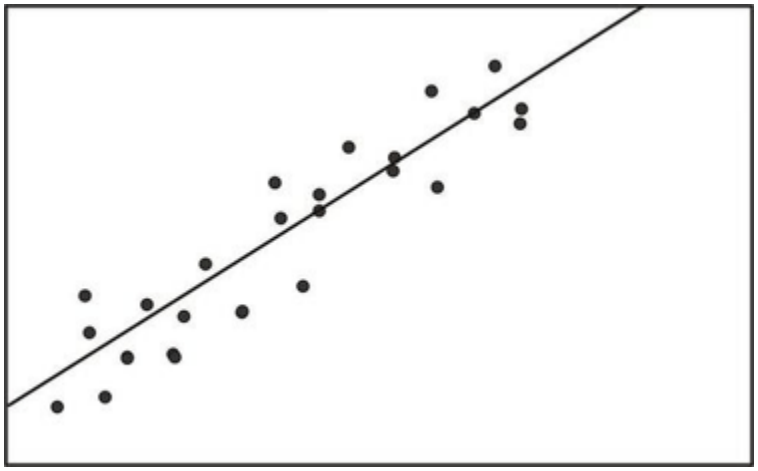
If the roles of the variables are not clear from context, you must decide which variable you want to predict. That variable becomes the response variable.

Example 4.2.2

Suppose we record SAT Math and SAT Verbal scores for a group of students. There is no inherent reason to treat one score as explaining the other if we are only interested in the relationship between them. However, if the purpose of the study is to predict SAT Math scores from SAT Verbal scores, then the Verbal score becomes the explanatory variable.

Interpreting Linear Regression Models

To develop a prediction model, we first need a strong linear relationship between the explanatory and response variables. When the correlation is strong, the points in a scatterplot tend to follow a straight-line pattern. If we could draw a line that best represents this pattern and write an equation for it, we could use that equation to make predictions.



This line is called the **least-squares regression line**, and the equation that makes the line is called the **linear regression**

equation. Sometimes, this equation will be called a **regression model** or **prediction model.** In the next section, we will learn how to determine the equation of the line that best fits the data. For now, we focus on using an equation for prediction.

In algebra, a line is typically written as

$$y = mx + b.$$

In statistics, we use a slightly different notation:

$$\hat{y} = a + bx$$

where

- $\hat{y}$ is the **predicted value** of the response variable. The hat on y emphasizes that the model produces predicted values, not actual ones.
- x is the value of the explanatory variable.
- a is the **y-intercept**, the predicted value of $\hat{y}$ when $x = 0$.
- b is the **slope**, the predicted change in $\hat{y}$ for each one-unit increase in x.

This regression equation provides a mathematical model that allows us to estimate values of the response variable based on observed values of the explanatory variable.

Example 4.2.3

A civil engineer creates a model that predicts the amount of water flowing in a stream (measured in gallons per minute) under a bridge using daily rainfall (in inches) as the explanatory variable. The regression equation is

$$\hat{y} = 1.6 + 29x.$$

The variable x represents the explanatory variable. Here, x is the amount of rainfall. When we substitute a rainfall amount into the equation, the model produces a predicted stream flow $\hat{y}$.

- If it rains 1 inch, substitute $x = 1$: $\hat{y} = 1.6 + 29(1) = 30.6$ gallons per minute.
- If it rains 2 inches, substitute $x = 2$: $\hat{y} = 1.6 + 29(2) = 49.6$ gallons per minute.
- If it rains 0.5 inch, substitute $x = 0.5$: $\hat{y} = 1.6 + 29(0.5) = 16.1$ gallons per minute.

Interpreting the Intercept

The y-intercept, $a = 1.6$, represents the predicted stream flow when $x = 0$. This means that even with no rainfall, about 1.6 gallons are expected to flow under the bridge every minute. This value reflects a baseline or "initial" flow that occurs regardless of rainfall - perhaps due to groundwater, upstream sources, or natural seepage.

Interpreting the Slope

The slope, $b = 29$, tells us how much the predicted stream flow changes for each additional inch of rainfall. In this model, the flow is expected to increase by 29 gallons per minute for every 1-inch increase in rainfall. The slope captures the strength of the relationship between rainfall and stream flow and is central to making predictions.

Example 4.2.4

Jamie collects data on the amount of time her classmates spend playing video games and doing homework. She models the relationship using a linear regression equation, measured in hours. Instead of using x and y, she labels the variables directly:

$$\widehat{\text{homework time}} = 23.6 - 1.65(\text{video game time}).$$

a. Which variable did Jamie use as the explanatory variable?

Solution

The explanatory variable is the variable used to make the prediction. In this equation, *video game time* is the explanatory variable because it is used to predict *homework time*.

b. Predict the amount of homework time for a student who plays video games for 5 hours.

Solution

Substitute 5 for video game time:

$$\widehat{\text{homework}} \text{ time} = 23.6 - 1.65(5) = 15.35$$

A student who plays video games for 5 hours is predicted to spend 15.35 hours on homework.

c. Identify and interpret the y-intercept.

Solution

The y-intercept is 23.6 hours. This represents the predicted homework time for a student who does not play video games at all. In other words, when video game time is 0 hours, the model predicts 23.6 hours of homework.

d. Identify and interpret the slope.

Solution

The slope is –1.65. This means that for every additional hour spends playing video games, the predicted homework time decreases by 1.65 hours. Because the slope is negative, the model describes a negative linear relationship. As video game time increases, homework time tends to decrease.

Example 4.2.5

A guidance counselor researches how attendance in district high schools relates to graduation rates at those same schools. The regression equation is

$$\hat{y} = -67.6 + 1.693x,$$

where x is the attendance rate and $\hat{y}$ is the graduation rate, both measured in percent.

a. Predict the graduation rate for a high school with an attendance rate of 85%.

Solution

Substitute $x = 85$ into the regression equation:

$$\hat{y} = -67.6 + 1.693(85) \approx 76.3.$$

A school with an attendance rate of 85% is predicted to have a graduation rate of about 76.3%.

b. Identify and interpret the y-intercept.

Solution

The y-intercept is −67.6. This represents the predicted graduation rate when attendance is 0%. In this context, a graduation rate of −67.6% is not meaningful. Graduation rates cannot be negative, and no school has 0% attendance.

Even though the intercept has no practical interpretation, it is still an essential part of the regression equation because it ensures the line fits the observed data as accurately as possible.

c. Identify and interpret the slope.

Solution

The slope is 1.693. This means that for every 1-percentage-point increase in attendance, the predicted graduation rate increases by 1.693 percentage points. The slope describes how sensitive graduation rate is to changes in attendance and reflects the positive linear relationship between the two variables.

You Try It $4.2.B$

A nutrition researcher studies how the number of hours a student sleeps each night relates to the student's daily concentration score (measured on a 100-point scale). She creates a linear regression model and obtains the equation:

$$\widehat{\text{concentration}}\ \text{score} = 42.5 + 6.8(\text{hours of sleep}).$$

Interpret this model by filling in the blanks:

a. In this model, the response variable is _____.

b. The concentration score for a student who sleeps 7 hours per night, is _____.

c. The y-intercept is _____, which represents the predicted concentration score for a student who sleeps _____ hours.

d. For each additional hour of sleep, the predicted concentration score changes by _____ points.

Answer available at the end of this section.

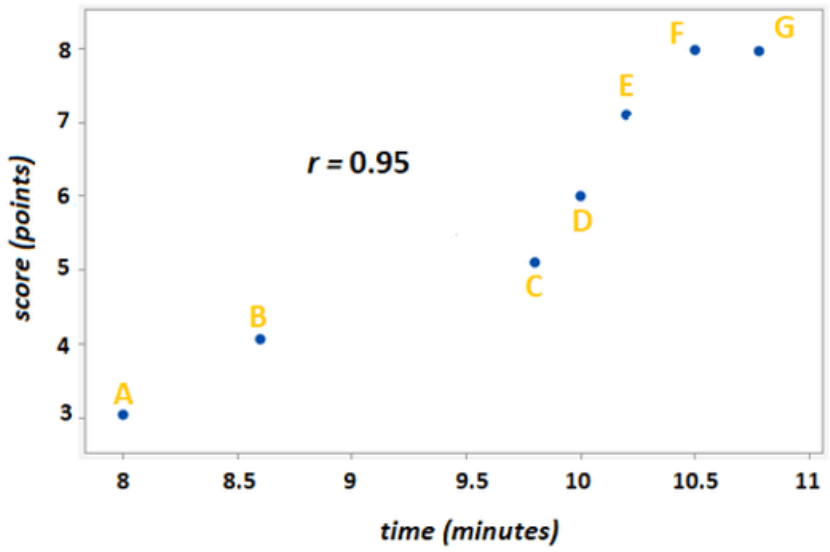
Developing Linear Regression Models

In this course, we will rely on technology to compute the slope and y-intercept of a least-squares regression line. Before learning how to calculate these values, it is important to understand conceptually what makes a regression equation "the best" representation of a linear trend in bivariate data.

A linear regression model is built from a sample of paired observations. The goal is to find values of a (the intercept) and b (the slope) so that the resulting line fits the scatterplot as closely as possible. Although there are many ways one might define "best fit," statisticians use a specific criterion based on minimizing prediction error.

Let's explore this idea of "best fit" with a scenario:

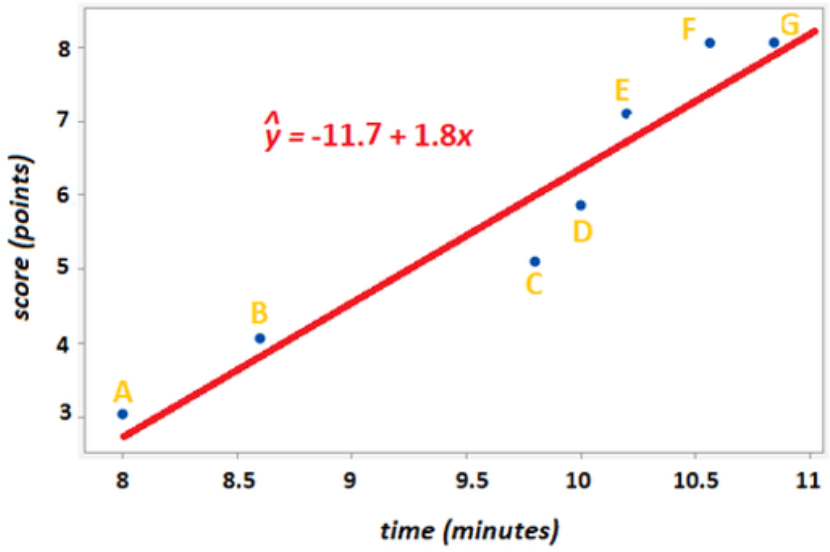
Mr. Thomas records how long each of his students spends completing a quiz (in minutes) and the corresponding quiz score. The scatterplot of seven students (A through G) shows a strong positive relationship, with a correlation of approximately $r \approx 0.95$.



Using this data, he constructs a regression model to predict a student's score from the time spent on the quiz. *Time* is the explanatory variable, and *score* is the response variable. The regression equation is

$$\hat{y} = -11.7 + 1.8x$$

and has been graphed on the scatterplot to show the overall trend.



Because the data are not perfectly linear, predictions made by the regression line will differ slightly from the actual scores. These differences are called **residuals**.

Understanding Residuals

A residual measures how far an observed data point is from its predicted value:

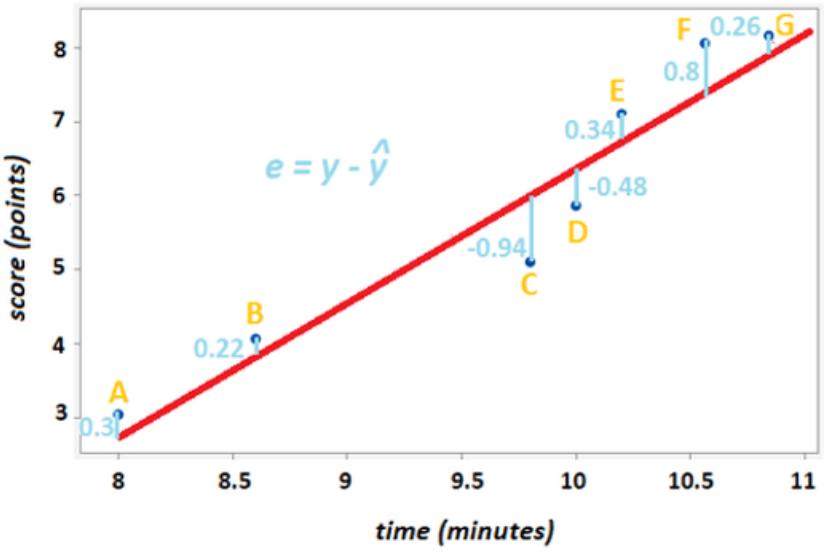
$$e = y - \hat{y}.$$

Student A spent $x = 8$ minutes on the quiz and scored $y = 3$ points. The model predicts a score of $\hat{y} = -11.7 + 1.8(8) = 2.7$ points.

The residual is $e = y - \hat{y} = 3 - 2.7 = 0.3$ points. This difference is positive. The model **under-predicted** the score because the predicted score $(\hat{y} = 2.7)$ was less than Student A's true score $(y = 3)$.

Student C spent $x = 9.8$ minutes on the quiz and scored $y = 5$ points. The model predicts a score of $\hat{y} = -11.7 + 1.8(9.8) = 5.94$ points. The residual is $e = y - \hat{y} = 5 - 5.94 = -0.94$ points. This difference is negative. The model **over-predicted** the score because the predicted score $(\hat{y} = 5.94)$ was greater than Student C's true score $(y = 5)$.

Residuals can also be viewed as vertical distances between the points and the regression line on the scatterplot. Points above the line have positive residuals; points below the line have negative residuals.



Why This Line Is “Best”

Statisticians define the **line of best fit** as the line that minimizes the sum of the squared residuals:

$$\dsum e^2 = \dsum (y - \hat{y})^2.$$

Squaring each residual emphasizes larger errors and ensures all contributions are positive. When these squared distances are added together, the least-squares regression line is the unique line that makes this total as small as possible. No other line produces a smaller sum of squared residuals.

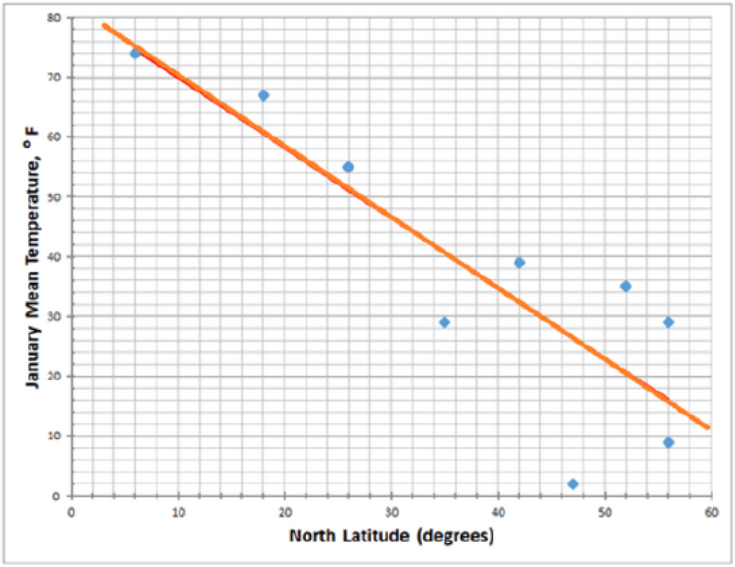
This criterion -- minimizing $\sum (y - \hat{y})^2$ -- is why the method is called **least-squares regression**. It provides a consistent, mathematically justified way to choose the single line that best represents the linear trend in the data.

Example 4.2.6

For her science fair project, Mina investigates how distance from the equator relates to winter temperatures. She records the north latitude and mean January temperature for nine cities in the Northern Hemisphere, then uses the data to construct a least-squares regression model. Her data table, scatterplot, correlation coefficient, regression equation, and graph are shown below.

Latitude and Temperature for Cities

City	Latitude $(^{o}\mathrm{N})$	January Mean Temperature $(^{o}\mathrm{F})$
Lagos, Nigeria	6	74
San Juan, Puerto Rico	18	67
Calcutta, India	26	55
Tokyo, Japan	35	29
Rome, Italy	42	39
Quebec City, Canada	47	2
London, England	52	35
Copenhagen, Denmark	56	29
Moscow, Russia	56	9



$$r \approx 0.86 \; \hat{y} = 81.79 - 1.175x$$

a. Which variable did Mina use as the explanatory variable?

Solution

Because *north latitude* appears on the x-axis, it serves as the explanatory variable, and *mean January temperature* is the response variable.

b. Predict the mean January temperature for a city located at latitude 50°N.

Solution

Substitute $x = 50$ into the regression equation:

$$\hat{y} = 81.79 - 1.175(50) = 23.04.$$

From either the equation or the graph, a city at 50°N is predicted to have a mean January temperature of about 23°F.

c. Identify and interpret the slope in context.

Solution

The slope is –1.175. This means that for each 1° increase in latitude, the predicted mean January temperature decreases by 1.175°F. The negative slope reflects a negative linear relationship: higher latitudes tend to be colder.

d. Identify and interpret the y-intercept in context.

Solution

The y-intercept is 81.79°F. This represents the predicted mean January temperature for a location at 0°N, the equator. In this context, the intercept is meaningful: it estimates the average January temperature at the equator based on Mina's model.

e. Does the model over-predict or under-predict the temperature for Tokyo, Japan?

Solution

Tokyo's data point is (35, 29). On the graph, the regression line lies above this point, meaning the model predicts a temperature higher than the observed value. Therefore, the model over-predicts Tokyo's mean January temperature.

f. Without calculating, is the residual for San Juan, Puerto Rico positive, negative, or zero?

Solution

San Juan's data point is (18, 67). Because the regression line is below this point, the observed temperature is greater than the predicted temperature. Thus, the residual is positive: $e = y - \hat{y} > 0$.

g. Which city appears to have a residual closest to 0? Furthest from 0?

Solution

A residual of 0 occurs when a point lies exactly on the regression line. The point closest to the line is (6, 74), corresponding to Lagos, Nigeria, so Lagos has the residual closest to zero. The point farthest from the line is (47, 2), corresponding to Quebec City, Canada, giving it the residual farthest from zero.

Section Summary

An explanatory variable is a variable used to help predict or explain changes in another variable, called the response variable. In a scatterplot, the explanatory variable is placed on the x-axis and the response variable on the y-axis.

Prediction involves estimating the value of one variable based on the value of another. When two quantitative variables show a strong linear relationship, we model that relationship using the least-squares regression line, written as a regression equation:

$$\hat{y} = a + bx.$$

The hat on y emphasizes that the model produces predicted, not actual, values. In this equation, a is the y-intercept, the predicted value of the response variable when $x = 0$, and b is the slope, the predicted change in the response variable for each one-unit increase in the explanatory variable.

To evaluate how well the model fits the data, we compare predicted values to actual observations. Sometimes the model may over-predict or under-predict a true value. A residual is the difference

$$e = y - \hat{y},$$

and represents the vertical distance between a data point and the regression line. Residuals may be positive or negative depending on whether the model under-predicts or over-predicts the observed value.

The least-squares regression line is the line that minimizes the sum of the squared residuals, $\sum (y - \hat{y})^2$, making it the single line that best fits the overall pattern of the data.

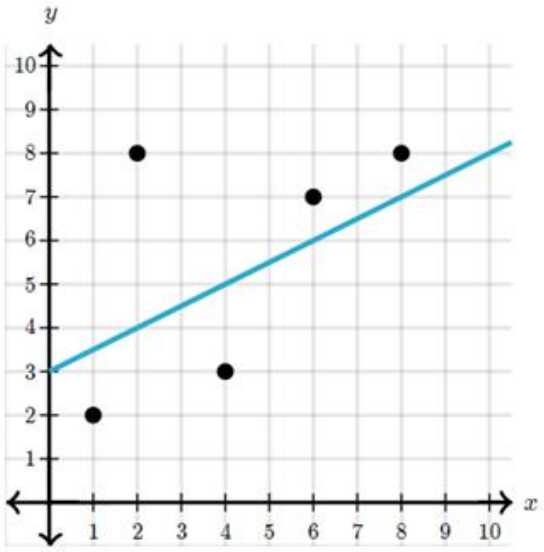
Section 4.2 Review Questions

1. In each study identify which variable is the explanatory variable. If it is not obvious that one variable is the explanatory variable, state '*not possible*.'
   a. The amount of time studying for a statistics exam and the grade on the exam.
   b. The yield of corn in bushels per acre and the amount of rain in the growing season.
   c. The number of texts and the number of calls for cell-phone users.
   d. The weight (in carats) and the price (in dollars) for a sample of diamonds.

2. The linear regression equation $\hat{y} = 51.75 + 1.65x$ models the average community college tuition per credit in a certain state for the past two decades. In the equation, x represents the number of years since 2001 and y is the tuition in dollars per credit.
   a. Which variable is the explanatory and which is the response?
   b. Predict the cost of tuition in the year 2015.
   c. Identify and interpret the value of the y-intercept in the context of this scenario. If the interpretation is not meaningful, state so.
   d. Identify and interpret the value of the slope in the context of this scenario.

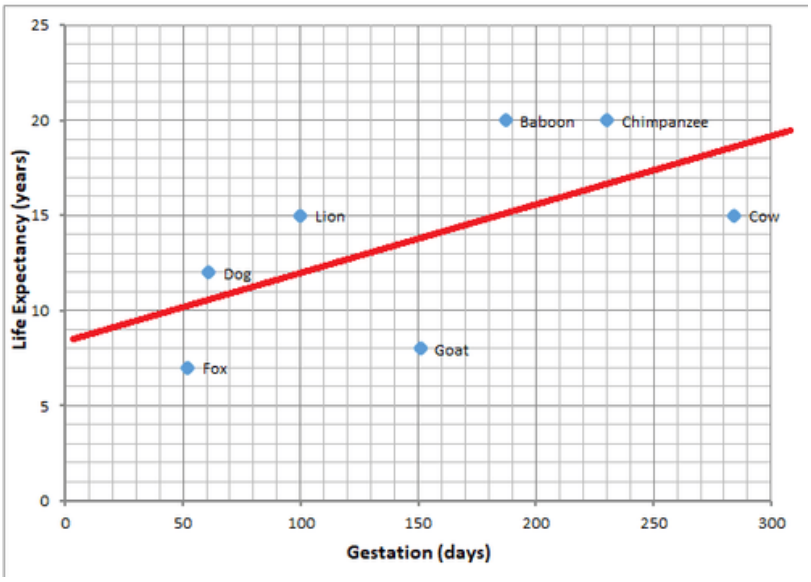
3. The price of a certain make and model of used car and the number of miles shown on its odometer have a strong correlation. The relationship between the variables can be described by the regression equation $\widehat{price} = 28400 - 0.16(\text{miles})$.
   a. Which variable is the explanatory and which is the response?
   b. Predict the price of the car if the car has been driven 5,000 miles.
   c. Identify and interpret the value of the y-intercept in context of the scenario. If the interpretation is not realistic, state so.
   d. Identify and interpret the value of the slope in context of the scenario.

4. A scatterplot of a simple set of data is shown along with its least squares regression line. For each point of the data, find y, $\hat{y}$, and the corresponding residual e.



5. Use the regression equation and scatterplot that relates *gestation period* (in days) and *life expectancy* (in years) for some animals to answer the questions.

$$\widehat{\text{life expectancy}} = 8.38 + 0.036(\text{gestation period})$$



a. Which variable is the response variable?
b. Identify and interpret the value of the slope.
c. Does this linear model under-predict or over-predict the cow's life expectancy?
d. Is the residual for the lion positive or negative?
e. Which animal has a residual value closest to 0?
f. Which animal has a residual value furthest from 0?
g. Calculate the residual for the fox.

Answers

You Try It $4.2.A$

It is reasonable to think that sunlight influences growth. Thus, *sunlight* is the explanatory variable x, and *plant growth* is the response variable y.

You Try It $4.2.B$

a. concentration score
b. 89.1
c. 42.5; 0
d. 6.8

This page titled 4.2: Interpreting a Least Squares Regression Model is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

4.3: Building a Least Squares Model

Learning Objectives

- Use technology to write the equation of a least squares regression line for a set of bivariate data.
- Use a least squares regression line to solve problems.
- Distinguish between interpolation and extrapolation in making predictions.

Introduction

In Section 4.2, we focused on interpreting a least-squares regression model once it had already been created. These models are built from sample data, and although no linear equation will fit every point in a scatterplot perfectly, a strong correlation allows us to use the model to make reasonable predictions.

In this section, we shift from *using* a regression model to constructing such a model with technology. By examining the structure of regression more closely, we will see how sample data lead to a specific equation and how that equation captures the overall trend in the relationship between two quantitative variables.

Using Technology to Find the Slope and y-Intercept

We now focus on writing the equation of the least-squares regression line from a set of data. Our goal is to determine the values of the y-intercept a and slope b in the model

$$\hat{y} = a + bx,$$

so that the sum of the squared residuals is minimized. This produces the line that best represents the linear trend in the data.

Although the slope and y-intercept can be computed manually, the algebra involved is beyond the scope of this course. Instead, we will rely on technology to generate the regression equation and then use that equation to interpret relationships, make predictions, and solve problems.

Example 4.3.1

Write the least squares regression equation for the line that best fits this simple set of bivariate data.

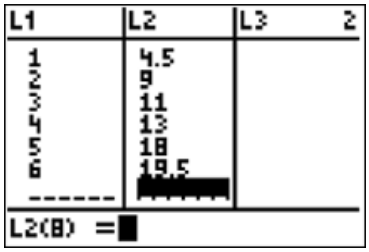
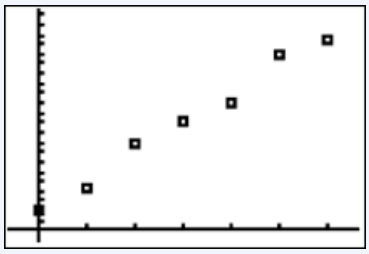
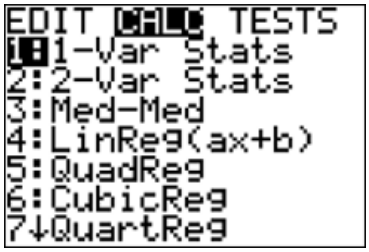
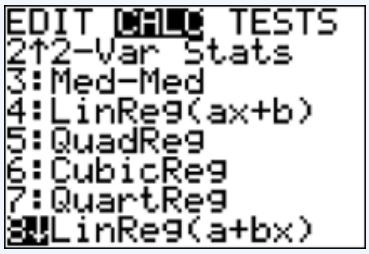
A set of bivariate data

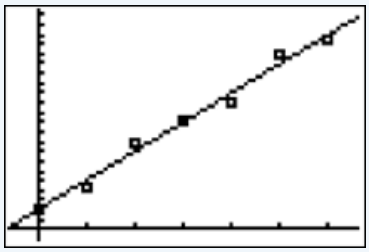
x	0	1	2	3	4	5	6
y	2	4.5	9	11	13	18	19.5

Solution

We will find the y-intercept and slope with a graphing calculator using the steps below.

Constructing a Regression Line on the TI-83/84 Calculator

1.	Press **[STAT]** and choose **1:Edit** to type the *x*-values in L1 and the *y*-values in L2.	
2.	Check to make sure the data are strongly correlated by displaying the data in a scatterplot. Press **[2nd] [Y=]** to access **STAT PLOTS**. Modify the options to show a scatterplot for the data as we have done before. Press **[ZOOM]** and **9:ZoomStat** to automatically fit the window to the data. Return to Section 4.1 to review the steps of making a scatterplot if needed.	
3.	Press **[STAT]** and use the right arrow to move to the **CALC** menu.	
4.	Choose **8:LinReg(a+bx).** This is the command we used previously to find the correlation coefficient. Note that **4:LinReg(ax+b)** looks similar, but the equation is not the form we use in statistics.	

5.	You will have one of the two screens shown here. If you have the top one, the blinking cursor is waiting for you to enter the names of the lists of data separated by a comma: L1,L2. Press **[ENTER]**. If you have the STAT WIZARD screen like the one on the bottom, choose L1 for **Xlist** and L2 for **Ylist**. You can leave the other options blank for now. Highlight **Calculate** and press **[ENTER]**.	 
6.	The value a is the y-intercept and the value b is the slope to match the statistical form of the line at the top. You can also see the correlation coefficient r. The linear regression equation is $\hat{y} \approx 2.054 + 2.982x$.	
7.	To check how well this regression line fits the data, press **[Y =]** and type the regression equation into equation **Y1**.	
8.	Press **[GRAPH]** to evaluate the fit of the linear model. Press **[TRACE]** to move to any data value so you can compare its value to the predicted value, $\hat{y}$, shown by the line.	
	Note: Instead of manually typing the regression equation in **[Y =]**, you can paste it there. While in the **[Y=]** screen, press **[VARS]**. Choose **5:Statistics**. Arrow over to **EQ** and select **1:RegEQ**.	

Example 4.3.2

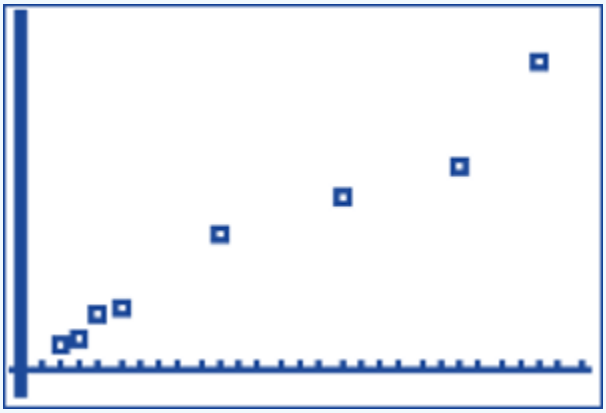
Peter is testing the burning time of Brite-n-Glo candles. The table shows how long candles of different weights burn. He wants to know whether a linear model is appropriate and, if so, how to interpret the resulting regression equation.

Candle Weight and Burn Time Data

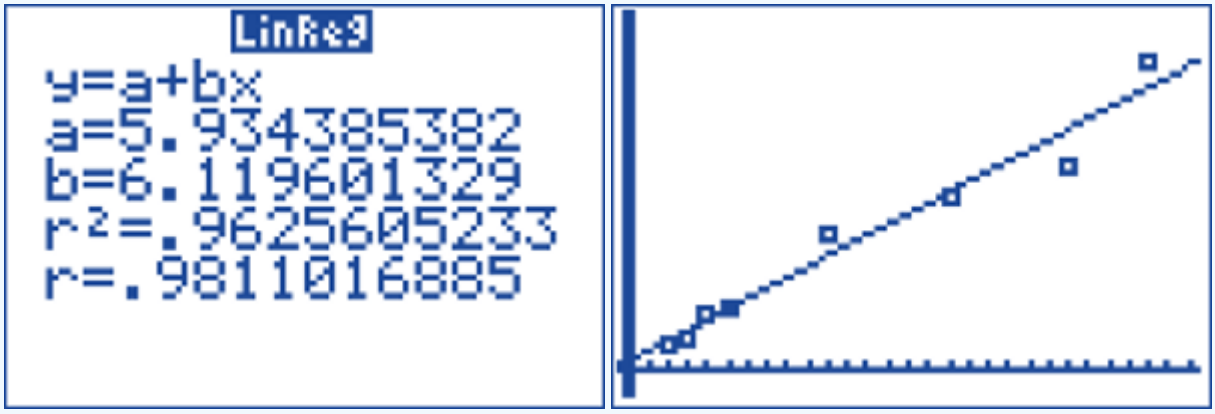
Candle weight (oz.)	Time (hours)
2	15
3	20
4	35
5	36
10	80
16	100
22	120
26	180

Solution

The explanatory variable is *candle weight*, and the response variable is *burning time*. A scatterplot of the data shows a strong linear trend with only minor scatter, so it is reasonable to model the relationship using a least-squares regression line.



Using a graphing calculator, the regression results and a graph of the regression line are as follows:



Peter can write the regression equation as

$$\widehat{\text{time}} \approx 5.93 + 6.12(\text{candle weight}).$$

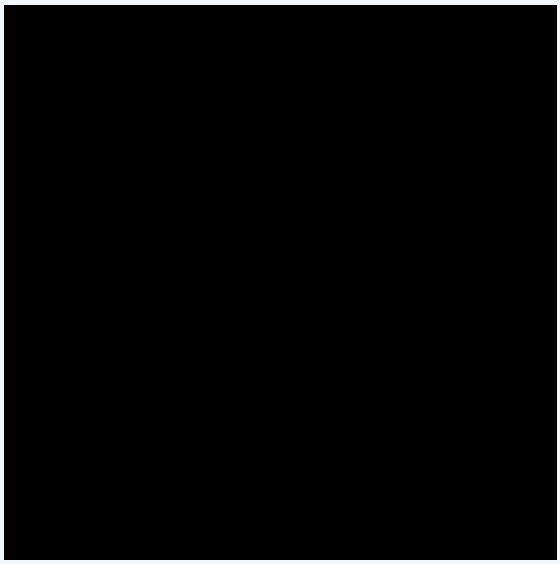
The regression line fits the data well because the points lie close to the line.

The slope of the regression equation is 6.12. This means that, on average, each additional ounce of candle weight increases the burning time by about 6.12 hours.

The y-intercept is 5.93. This represents the predicted burning time for a candle weighing 0 ounces. Because a candle cannot weigh zero ounces, this value does not have a practical interpretation, but it is still necessary for defining the regression line mathematically.

Example 4.3.3

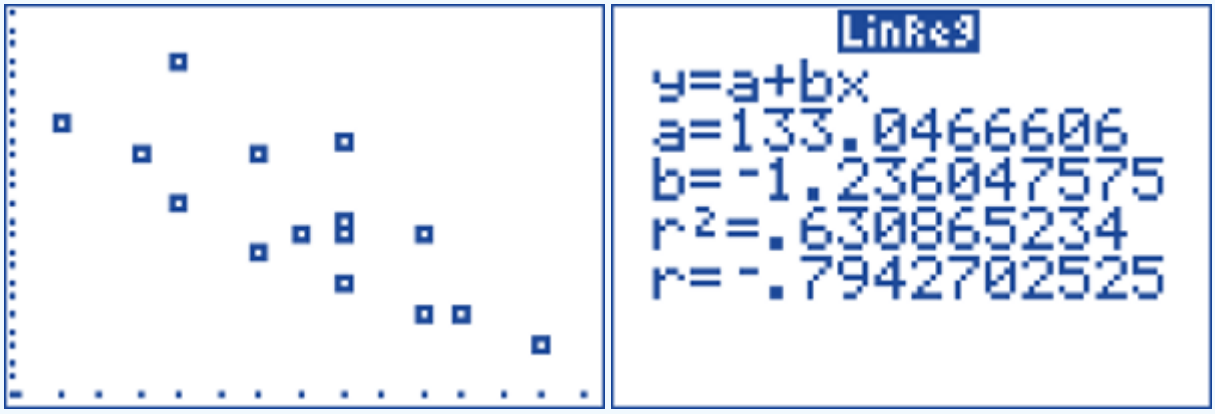
Ms. Fox collected data from her high school class by measuring each student's height and then recording how many steps each student took to walk the length of a school hallway.



a. Predict the number of steps for a student who is 65 inches tall.

Solution

Using a graphing calculator, we treat *height* as the explanatory variable and *number of steps* as the response variable.



The regression equation is $\widehat{\text{steps}} = 133.05 - 1.24(\text{height})$.

Substituting 65 for height, $\widehat{\text{steps}} = 133.05 - 1.24(65) \approx 52.05$.

A student who is 65 inches tall is predicted to take about 52 steps to walk the hallway.

b. What does the slope tell you about height and number of steps?

Solution

The slope is –1.24. This means that for every additional inch of height, a student is predicted to take 1.24 fewer steps. Taller students generally take longer strides, so they need fewer steps to cover the same distance.

c. Find the residual for Student A.

Solution

Student A is 64 inches tall and took 51 steps. Substituting their height (64) into the regression equation, we can predict that Student A's number of steps as

$\widehat{\text{steps}} = 133.05 - 1.24(64) \approx 54$ steps.

The residual is

$$e = \text{steps}(y) - \widehat{\text{steps}}(\hat{y}) = 51 - 54 = -3.$$

The residual is negative, meaning the model over-predicts Student A's number of steps by 3.

d. Another student (not in the original data) took 47 steps. How tall do you think this student is?

Solution

To estimate the height, solve the regression equation for the value of height x that predicts $\hat{y} = 47$ steps:

$$\begin{aligned} 47 &= 133.05 - 1.24x \\ 1.24x &= 86.05 \\ x &= \frac{86.05}{1.24} \\ x &= 69.4 \end{aligned}$$

A student who takes 47 steps is predicted to be 69.4 inches tall.

Interpolation and Extrapolation

Interpolation and extrapolation are two ways to use a regression model to predict hypothetical values of a response variable based on an explanatory variable.

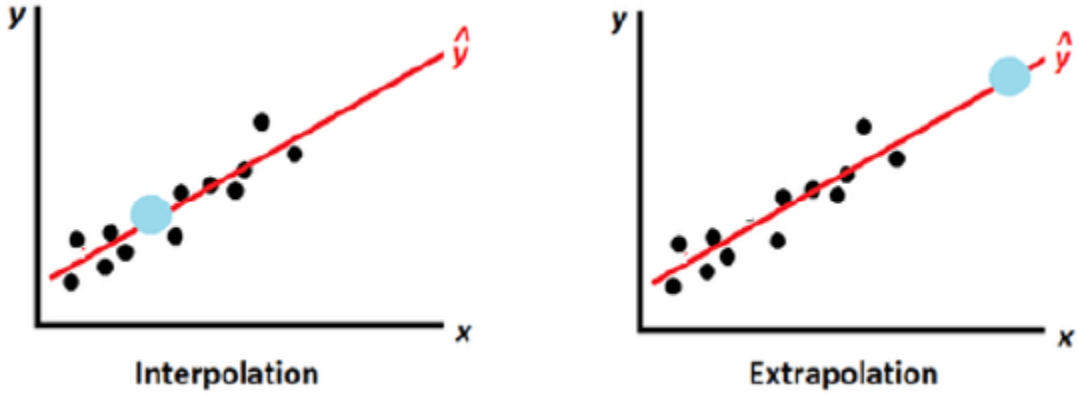
Interpolation occurs when we make a prediction using an explanatory value within the range of the data used to create the regression line. Because the model is being applied to familiar territory—values similar to those already observed—interpolated predictions are generally more reliable.

In Example 4.3.3, we predicted the number of steps for a student who is 65 inches tall. This is interpolation because $x = 65$ lies between the observed heights of 61 inches and 73 inches. We are fairly confident of the trend between heights of 61 and 73 inches.

Extrapolation, on the other hand, occurs when we make a prediction using an explanatory value outside the range of the original data. These predictions are riskier because they assume that the trend observed in the sample continues beyond the values we actually measured. We don't know that they actually do.

For instance, using the model from Example 4.3.3 to predict the number of steps for a 40-inch-tall 3rd grader is extrapolation. The value $x = 40$ is far below the heights of the sampled high school students [61,73] that were used to write the regression equation. The model predicts $\widehat{\text{steps}} = 133.05 - 1.24(40) \approx 83$ steps, but we cannot be as confident that this is a good prediction.

The figure below illustrates this distinction clearly: interpolation stays within the data range, while extrapolation extends beyond it.



Interpolation is preferred because it is more likely to produce meaningful estimates. Extrapolation requires caution; the relationship between variables may change outside the observed range, and predictions can become unrealistic or misleading. When a model produces such unreasonable results, we call this an extrapolation error.

Section Summary

In this part of the chapter, we moved from interpreting a regression model to understanding how regression equations are created and evaluated. A least-squares regression line models the linear relationship between two quantitative variables by choosing the slope and y-intercept that minimize $\sum (y - \hat{y})^2$, the sum of the squared residuals. Although the calculations behind this process are algebraically complex, technology allows us to generate the regression equation efficiently and use it to make predictions and solve problems.

We also distinguished between interpolation and extrapolation when using a regression model for prediction. Interpolation uses explanatory values within the range of the original data and typically produces reliable estimates. Extrapolation uses values outside that range and must be approached cautiously, as the observed trend may not continue beyond the data. Extreme extrapolation can lead to unrealistic or meaningless predictions, known as extrapolation errors.

Together, these ideas form the foundation for building, interpreting, and responsibly using linear regression models to analyze real-world relationships.

Section 4.3 Review Questions

1. The table shows the paired SAT math scores and GPAs for 7 students.

Problem 1: SAT Math Scores and GPAs

SAT math score, x	595	520	715	405	680	490	565
GPA, y	3.4	3.2	3.9	2.3	3.9	2.5	3.5

   a. Use technology to make a scatterplot of the data. Use SAT math score as the explanatory variable. Visually, does there appear to be a strong correlation?
   b. Use a calculator to write the least squares regression equation. *Round all values to four decimal places.*
   c. What is the value of the correlation coefficient? Explain how its value confirms your answer to part (a).
   d. Use the equation to predict the GPA for a student with an SAT score of 500. *Round GPA to the nearest hundredth*. Did you interpolate or extrapolate to make the prediction?
   e. Draw and complete a table like the one below. Fill in the predicted values and residuals for each student. *Round each GPA to the nearest hundredth.*

Problem 1e: SAT Math Scores, GPAs, Predicted GPAs and Residuals

SAT math score, x	595	520	715	405	680	490	565
GPA, y	3.4	3.2	3.9	2.3	3.9	2.5	3.5
predicted GPA, $\hat{y}$							
residuals, $y - \hat{y}$							

2. A school nurse wants to predict scores on a memory test from the number of times that a student exercises per week. Below are her observations from several students. Use technology to answer the questions below.

Problem 2: Memory Test Scores vs. Exercise Frequency

Student #	Exercise (days per week)	Memory Test Score
1	0	15
2	2	3
3	2	7

Student #	Exercise (days per week)	Memory Test Score
4	1	11
5	3	5
6	1	8
7	2	10
8	0	13
9	3	2
10	3	4
11	4	2
12	1	8
13	1	10
14	1	12
15	2	8

a. Use technology to make a scatterplot. Be sure choose the correct variables for the explanatory and response variables.
b. Does this appear to be a linear relationship? Why or why not?
c. Find the least squares regression equation used to model this data. *Round all values to four decimal places.*
d. Predict the memory test score of a student who exercises 3 times per week. *Round the score to the nearest whole number.*
e. What is the slope in this linear regression model, and what does this mean in this scenario?
f. What is the *y*-intercept in this linear regression model? If it has meaning in this scenario, interpret its value.
g. Draw the regression line on the scatterplot. Identify the student for which the model gives a residual value furthest from 0.
h. Find the residual for Student #11.
i. Predict the memory test score of a student who exercises every day of the week. Do you have concerns about making this prediction? Why or why not?

3. James and Keisha kept track of the temperature and the number of cups of coffee sold over a 9-day period at their food

truck. Their data are shown below.

Problem 3: Temperature and Coffee Data

Temperature ^{o}F	9	18	20	26	28	31	32	36	48
Cups of Coffee	45	41	34	36	32	28	23	23	18

c. Use technology to construct a scatterplot. Does there appear to be a positive, negative, or no correlation between the temperature and the number of cups of coffee sold?
d. Based on the scatterplot, do there appear to be any outliers? If so, which points?
e. Use technology to find the correlation coefficient.
f. Use technology to write the regression equation to predict the number of cups of coffee sold. *Round all values to two decimal places.*
g. Use the regression line from part d to predict the amount of coffee sold if the temperature is 22 degrees.
h. Keisha would like to use the regression line to predict the amount of coffee sold if the temperature is -10 degrees. Is this an appropriate use of the regression line? Why or why not?
i. Identify and interpret the value of the slope in this scenario.
j. Find the residual for the cups of coffee on the day when the temperature was 48 degrees.

4. Anne is trying to find the elasticity coefficient of a Superball in a science experiment. She drops the ball from different heights and measures the maximum height of the ball after the bounce. The table below shows the data she collected.

Problem 4: Superball Science Experiment

Initial height (cm)	**Bounce height (cm)**
30	22
35	26
40	29
45	34
50	38
55	40
60	45
65	50

Initial height (cm)	Bounce height (cm)
70	52

a. Use technology to draw a scatter plot using initial height of the ball as the explanatory variable. Is there a strong correlation?
b. Create the regression equation to predict bounce height. *Round values to places*.
c. Identify and interpret the value of the slope in this scenario.
d. Identify the value of the y-intercept. Interpret its value if there is meaning.
e. Anne wants the ball to bounce 65 cm. From what height should she drop the ball? Do you have concerns about whether this is a reasonable answer? Why or why not?

This page titled 4.3: Building a Least Squares Model is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

5: AN INTRODUCTION TO PROBABILITY

5.1: Theoretical Probability
5.2: Empirical Probability and the Law of Large Numbers
5.3: Using Sets and Venn Diagrams to Represent Events
5.4: Contingency Tables, Conditional Probability, and Independence
5.5: Odds

This page titled 5: An Introduction to Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

5.1: Theoretical Probability

Learning Objectives

- Use foundational probability terminology.
- List outcomes for events and sample spaces.
- Apply theoretical probability.

Introduction

Uncertainty is part of everyday life, and we constantly make decisions based on what we think is likely to happen. Imagine you are planning a picnic and notice that it has rained three afternoons in a row. You might wonder whether it is likely to rain again today. Or consider flipping a coin: if it has landed on heads four times in a row, you might question whether the next flip will finally be tails—or whether something unusual is going on with the coin.

These situations feel familiar because they involve judging the chance of an event. We look at patterns, compare past outcomes, and make predictions, even when we cannot be certain. Rolling the same number several times or seeing repeated weather patterns may seem surprising, and those surprises often lead us to rethink what we expect next.

In this chapter, you will explore mathematical ideas that help make sense of such everyday uncertainties. By learning how to describe random events, analyze outcomes, and interpret patterns, you will develop tools that support clear and confident reasoning. These ideas will not only deepen your own understanding of probability but also prepare you to guide future students as they encounter randomness in games, experiments, and real-world situations.

Experiments and Outcomes

In probability, an **experiment** is any process that produces a result determined by chance. An experiment may involve taking a measurement, making an observation, or performing a simple action. Flipping a coin, rolling a die, drawing a card, or observing tomorrow's weather are all examples of experiments arising from random phenomena.

An **outcome** is any result that can occur in an experiment. For the experiment of flipping a coin, the possible outcomes are *heads* or *tails*. For rolling a regular six-sided die, the outcomes are the numbers 1 through 6. Outcomes represent the most basic results we can observe. A list of all possible simple outcomes of the experiment is called the **sample space** of the experiment.

An **event** is a collection of outcomes. Some events contain only one outcome; these are called **simple events**. Other events contain more than one outcome; these are called **compound events**. Events allow us to group outcomes in meaningful ways.

Example 5.1.1

A standard six-sided die has six possible outcomes: $\{1, 2, 3, 4, 5, 6\}$. An experiment consists of rolling the die once. When we roll it, exactly one of these outcomes occurs. Therefore, the sample space for this experiment is $S = \{1, 2, 3, 4, 5, 6\}$. The **cardinality** of the sample space, written $n(S)$, is the number of outcomes in the set. Here, $n(S) = 6$.

- Consider the simple event A: rolling a 2 on a die. This is called a **simple event** because it contains only one outcome: $\{2\}$. We say the cardinality of this event is 1 and can write $n(A) = 1$ because there is only one outcome in event A.
- Consider the compound event B: rolling an odd number on a die. This is called a **compound event** because it includes several outcomes: $\{1, 3, 5\}$. The cardinality of this event is 3 and we can write $n(B)$ because there are three simple outcomes in this compound event.

These examples illustrate how events help us describe sets of outcomes that share a common feature as well as their cardinalities.

Example 5.1.2

John is playing a game with a spinner divided into four equal sections labeled 1, 4, 5, and 8. John must spin the spinner once.

a. Write the sample space S and find its cardinality.

Solution

$S = \{1, 4, 5, 8\}$ and $n(S) = 4$.

b. Event A is defined as getting an even number on the spin. Write the outcomes in Event A and find the cardinality of Event A.

Solution

$A = \{4, 8\}$ and $n(A) = 2$.

You Try It 5.1.A

A bag contains three colored cubes: one red (R), one blue (B), and one green (G).

For an experiment, you reach into the bag, mix the cubes around without looking, and draw one cube.

a. List the sample space S and find its cardinality.
b. Event E is defined as "*drawing a cube that is not red.*" Write the outcomes in Event E and find the cardinality of Event E.

Answer available at the end of this section.

Outcomes in Multi-Step Experiments

When an experiment involves more than one step, the outcomes in its sample space become more detailed. Instead of a single result, such as one coin flip or one die roll, each outcome must describe the result of *every* step in the experiment. For example, if two coins are tossed, an outcome must show how both coins land. If two balls are drawn from a box, an outcome must show the color drawn first and the color drawn second. Multi-step experiments therefore produce sample spaces whose outcomes are combinations of simpler results, and understanding how to list these outcomes is essential for building accurate sample spaces.

Example 5.1.3

An experiment consists of tossing a penny and a quarter simultaneously. Each coin can land heads (H) or tails (T).

Writing the result of the penny followed by the quarter, the sample space is

$$S = \{HH, HT, TH, TT\}.$$

Here, $n(S) = 4$.

Now, define event A as "*getting at least one heads*." The outcomes in event A are

$$A = \{HH, HT, TH\}.$$

Since there are three outcomes in this event, we have $n(A) = 3$.

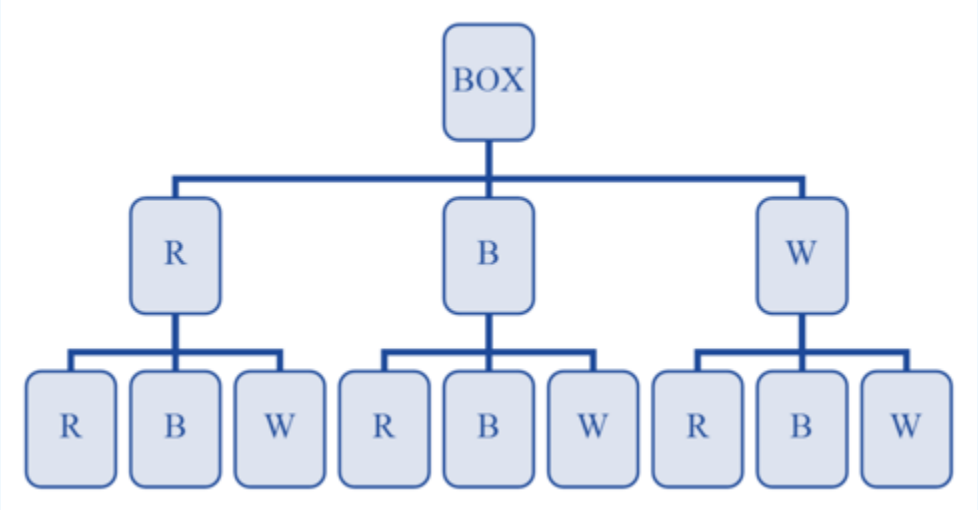
Example 5.1.4

Suppose a box contains three balls—one red (R), one blue (B), and one white (W). One ball is selected, its color is observed, and then the ball is placed back into the box. After mixing the balls again, a second ball is selected and its color is observed.

List the sample space for this experiment and find its cardinality.

Solution

A tree diagram helps illustrate all possible selections:



For the first selection, the result could be R, B, or W. Because the ball is replaced in the box, the second selection could also be R, B, or W.

Each pair of letters represents the colors drawn on the first and second selections, respectively. For example, RR means red on both draws, and RW means red first and white second. The sample space and its cardinally is

$$S = \{RR, RB, RW, BR, BB, BW, WR, WB, WW\}\ n(S) = 9.$$

We can also represent this sample space in a table, where rows show the first draw and columns show the second draw. Each cell corresponds to one outcome and there are 9 cells in a 3 x 3 table.

First Draw \ Second Draw	R	B	W
R	RR	RB	RW
B	BR	BB	BW
W	WR	WB	WW

You Try It 5.1.B

Consider the same box of three balls as in Example 5.1.3. This time, after drawing one ball and observing its color, we do not place it back into the box. We then draw a second ball and record its color. Because the first ball is not replaced, the second draw has fewer possibilities.

List the sample space for this experiment and find its cardinality.

Answer available at the end of this section.

Theoretical Probability

Once we have identified the sample space and the events we care about, we can begin assigning probabilities. **Theoretical probability** describes how likely an event is to occur based on the structure of the experiment itself. It relies on the idea that each outcome in the sample space is **equally likely**. When this condition is met, probability becomes a simple comparison between how many outcomes make an event happen and how many outcomes are possible overall.

We use the notation $P(E)$ to represent the probability that event E occurs. If E is an event within a sample space S, then the theoretical probability of E is defined by the ratio:

$$P(E) = \frac{\text{number of outcomes in event } E}{\text{number of outcomes in the sample space } S}$$

Using cardinality notation, this becomes

$$P(E) = \frac{n(E)}{n(S)}$$

This formula applies only when each outcome in the sample space is equally likely to occur. Situations such as flipping a fair coin, rolling a fair die, or selecting a card from a well-shuffled deck satisfy this condition and allow us to compute theoretical probabilities reliably.

Example 5.1.5

When rolling a standard die, what is the probability of rolling a multiple of 3?

Solution

A standard die has 6 equally likely outcomes in the sample space: $S = \{1, 2, 3, 4, 5, 6\}$.

If Event E is "*getting a multiple of 3*" , then $E = \{3, 6\}$ and $n(E) = 2$.

$P(\text{rolling a multiple of } 3) = \frac{n(E)}{n(S)} = \frac{2}{6} = \frac{1}{3} \approx 33\%$.

Example 5.1.6

When rolling a standard die, what is the probability of rolling a 7?

Solution

If Event E is "*rolling a 7*" , then $E = \{\ \}$. There are no outcomes in this event so $n(E) = 0$.

$P(\text{rolling a } 7) = \dfrac{n(E)}{n(S)} = \dfrac{0}{6} = 0 = 0\%.$

Example 5.1.7

A jar contains five marbles: two green, one yellow, and two purple. You reach in, mix them around, and draw one marble at random. What is the probability that the marble you draw is green?

Solution

The sample space consists of 5 equally likely outcomes: $S = \{G_1, G_2, Y, P_1, P_2\}$.

The event "drawing a green marble" contains 2 outcomes, so $n(E) = 2$.

$P(\text{draw a green marble}) = \dfrac{n(E)}{n(S)} = \dfrac{2}{5} = 40\%.$

You Try It 5.1.C

A shelf contains eight books: three novels, two biographies, and three science books. You close your eyes, reach out, and select one book at random. What is the probability that the book you select is a biography?

Answer available at the end of this section.

Example 5.1.8

A bag contains small tiles, each printed with a letter from the word **MATHEMATICS**. You reach in and draw one tile.

What is the probability that the tile shows the letter **A**?

Solution

The sample space consists of the 11 tiles that are all equally likely to be selected so $n(S) = 11$. T

he event "drawing an A" contains 2 outcomes because there are two such tiles, so $n(E) = 2$.

$P(\text{draw an A}) = \frac{n(E)}{n(S)} = \frac{2}{11} \approx 0.18 \approx 18\%.$

Example 5.1.9

Using the same bag of tiles from Example 5.1.7, you reach in and draw one tile without looking at the result.

This time, you are told that the tile you drew is a vowel.

Given that information, what is the probability that the tile shows the letter **A**?

Solution

If you were drawing a tile normally, every letter in **MATHEMATICS** would be part of the sample space. But once you are told that the tile is a vowel, you no longer consider any consonants at all. They are not possible outcomes anymore.

So you restrict the sample space to only the four vowels in the word. The "given" information shrinks the sample space.

You are simply comparing the number of "A" tiles $n(E) = 2$ to this smaller group of vowels $n(S) = 4$, not to the entire word.

$P(\text{draw an A given it is a vowel}) = \frac{n(E)}{n(S)} = \frac{2}{4} = \frac{1}{2} = 0.5 = 50\%.$

You Try It 5.1.D

A box contains tiles showing the letters in the word **SUNSHINE**. You draw one tile at random.

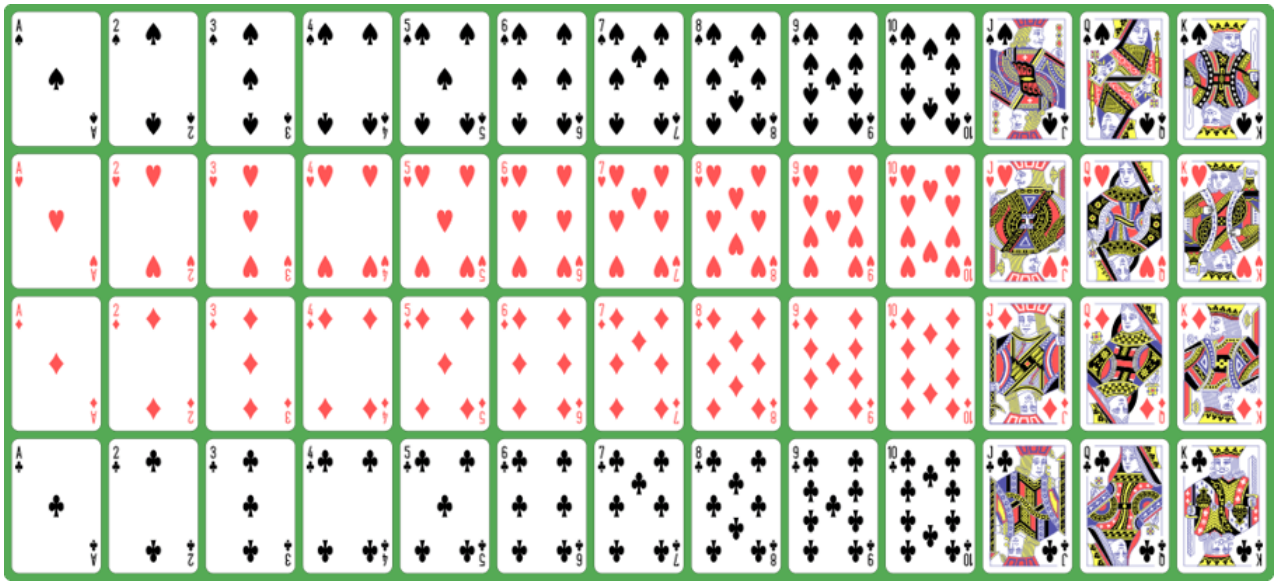
a. What is the probability that the tile shows the letter **N**?
b. Given the tile is a consonant, what is the probability the tile shows the letter **N**?

Answer available at the end of this section.

Working with a Standard Deck of Cards

You will often see probability problems framed in the context of drawing cards from a standard deck of playing cards. Because many students -- and some teachers -- may not be fully familiar with how such a deck is organized, it is helpful to understand its structure before working with card-based probability questions.

A standard deck contains 52 unique cards, all equally likely to be drawn. These cards are divided into four suits: hearts, diamonds, clubs, and spades. Hearts and diamonds are red; clubs and spades are black. Each suit contains 13 cards -- the numbers 2 through 10, along with Ace, Jack, Queen, and King. Here is a picture of a complete deck of cards for your reference in problems about cards.



Because the deck is structured and each card is distinct, it provides a clean and consistent sample space for many introductory probability problems. Teachers often use playing cards in the classroom because they are inexpensive, familiar to many students, and provide a rich set of outcomes for exploring ideas such as sample spaces, events, and equally likely outcomes. Understanding how the deck is designed will help you guide students confidently through card-based probability activities.

Example 5.1.10

Find the probability of drawing each of the following cards on the first draw from a standard deck of cards.

a. a red card

Solution

There are 52 cards in a deck so $n(S) = 52$. There are 26 red cards (13 hearts and 13 diamonds) so $n(E) = 26$.

$$P(\text{red card}) = \frac{n(E)}{n(S)} = \frac{26}{52} = \frac{1}{2}.$$

b. a club

Solution

There are 13 clubs in the deck so $n(E) = 13$.

$$P(\text{club}) = \frac{n(E)}{n(S)} = \frac{13}{52} = \frac{1}{4}.$$

c. a King

Solution

There are 4 Kings, one in each suit so $n(E) = 4$.

$$P(\text{King}) = \frac{n(E)}{n(S)} = \frac{4}{52} = \frac{1}{13}.$$

d. the King of clubs

Solution

There is exactly one King of clubs in the deck so $n(E) = 1$.

$$P(\text{King of clubs}) = \frac{n(E)}{n(S)} = \frac{1}{52} = \frac{1}{52}.$$

You Try It $5.1.E$

Find the probability of drawing a face card on the first draw from a standard deck of cards.

Answer available at the end of this section.

Example $5.1.11$

When tossing two coins, what is the theoretical probability of getting a head on both coins?

Solution

The sample space is $S = \{HH, HT, TH, TT\}$ so $n(S) = 4$.

The event "*getting a head on both coins*" is $E = \{HH\}$ and $n(E) = 1$.

$$P(2 \text{ heads}) = \frac{n(E)}{n(S)} = \frac{1}{4} = 0.25 = 25\%.$$

Example 5.1.12

The box contains three balls, one red (R), one blue (B), and one white (W). One ball is selected, its color is observed, and then the ball is placed back into the box. After mixing the balls again, a second ball is selected and its color is observed.

What is the probability that at least one of the two balls drawn is red?

Solution

From Example 5.1.4, the sample space for this experiment is

$$S = \{RR, RB, RW, BR, BB, BW, WR, WB, WW\}$$

and $n(S) = 9$.

The event "at least one red ball" includes all outcomes that contain an R in either position:

$E = \{RR, RB, RW, BR, WR\}$ and $n(E) = 5$.

$$P(\text{at least one red}) = \frac{n(E)}{n(S)} = \frac{5}{9} \approx 0.56 \approx 56\%.$$

You Try It 5.1.F

When tossing two coins, what is the theoretical probability of getting at least one tail?

Answer available at the end of this section.

Example 5.1.13

Two students, Sara and Ben, are participating in a quick "grab-a-token" challenge in their math class. Sara has a small bag containing the tokens $\{3, 8, 11\}$. Ben has a different bag containing the tokens $\{4, 7, 12\}$. Each student reaches into their own bag and draws one token at random. They compare their numbers.

a. List the sample space for one draw using ordered pairs (Sara, Ben).

Solution

Sara can draw 3, 8, or 11. Ben can draw 4, 7, or 12. The ordered pairs (Sara, Ben) are

$$S = \{(3,4), (3,7), (3,12), (8,4), (8,7), (8,12), (11,4), (11,7), (11,12)\}$$

and $n(S) = 9$.

b. What is the probability that Sara's number is larger than Ben's number?

Solution

Sara wins the challenge in these outcomes:

$$E = \{(8,4), (8,7), (11,4), (11,7), (11,12)\}$$

Here, $n(E) = 5$.

$$P(\text{Sara wins}) = \frac{n(E)}{n(S)} = \frac{5}{9}.$$

c. What is the probability that the sum of the two numbers is even?

Solution

The sum is even in these outcomes:

$E = \{(3,7), (11,7), (8,4), (8,12)\}$ and $n(E) = 4$.

$$P(\text{sum is even}) = \frac{n(E)}{n(S)} = \frac{4}{9}.$$

Section Summary

This section introduced the foundational ideas needed to describe and analyze probability experiments clearly and consistently. An experiment is any process that produces an outcome based on chance -- such as drawing a card, rolling a die, or selecting an object. An outcome is a result of that experiment and an event is collection of outcomes from that experiment. A simple event contains exactly one outcome, while a compound event contains two or more outcomes.

The sample space, usually denoted by S, is the complete set of all possible outcomes of an experiment. Once the sample space is identified, we can determine how many outcomes belong to any event E by counting its elements, written $n(E)$. Likewise, $n(S)$ represents the number of outcomes in the sample space.

When every outcome in the sample space is equally likely, we can compute the theoretical probability of an event using the ratio of the number of outcomes in the event to the number of outcomes in the sample space. In symbols,

$$P(E) = \frac{\text{number of outcomes in event } E}{\text{number of outcomes in the sample space } S} = \frac{n(E)}{n(S)}.$$

These ideas -- experiments, outcomes, events, sample spaces, and theoretical probability -- form the foundation for all probability reasoning you will encounter in the rest of this chapter.

Section 5.1 Review Questions

1. Consider an experiment of rolling a die and then flipping a coin. Write the sample space for this experiment. Then, use the sample space to find the probability of each event:
   a. roll a 2 on the die and get Heads on the coin
   b. roll an even number on the die and get Tails on the coin
   c. roll an even number on the die
   d. get Tails on the coin
   e. roll 7 on the die

2. A box contains ping pong balls. On each ping pong ball is a letter in the word PROBABILITY. You draw a single ball.
   a. What is the probability that you draw a consonant? (Consider Y a consonant.)
   b. What is the probability that you do not draw a “B”?
   c. Given that the ball is a consonant, what is the probability that it is “B”?

3. A box contains two blue marbles and one red marble. Two marbles are drawn randomly without replacement. Refer to the blue marbles as B1 and B2 and the red one as R. List the outcomes in the sample space for this experiment. Then, use the sample space to find the probability of each event below:
   a. draw two blue marbles
   b. draw one red marble and one blue marble, in any order
   c. draw a blue marble and then a red marble, in that order

4. Find the probability of drawing each on the first draw from a deck of playing cards:

a. draw a 3
b. draw a red 3
c. draw a face card (King, Queen, or Jack)

5. Jack and Jill are playing a game called War. In the game of War, Jack will draw a card from his deck and Jill will draw a card from her deck. They will then compare their cards. Whoever has the largest card wins the "battle."

   Jack's deck contains the cards 4, 7, and 9. Jill's deck contains the cards 2, 6, and 8.

   a. List the sample space for one battle of War using ordered pairs (Jack, Jill).
   b. What is the probability that Jack wins the battle?
   c. What is the probability that both cards are even?
   d. What is the probability that the sum of the two cards is a multiple of 3?
   e. What is the probability that the sum of the two cards is 15?

6. You will pick two friends to attend a movie with you. You want to choose one friend from school and one friend from church. Your school friends include Anthony, Bea, Cathy, and Darlene. Your church friends include Evelynn and Frank.
   a. List the complete sample space for the two friends that you could choose.
   b. What is the probability that Cathy will be chosen for the movie?
   c. What is the probability that both friends chosen will be female?
   d. What is the probability that the school friend is male and the church friend is female?
   e. What is the probability that one male and one female will be chosen (in any order)?
   f. What is the probability that at least one female will be chosen?

7. In an experiment, you will toss a fair coin three times and record whether it lands on heads or tails on each toss.

   a. List the sample space using H for heads and T for tails. *Hint:* There are 8 equally likely outcomes in the sample space.
   b. What is the probability that all three tosses will end up as heads?
   c. What is the probability that exactly one of the three tosses will end up as heads?
   d. What is the probability that at least one of the three tosses will end up as heads?
   e. What is the probability that at most one of the three tosses will end up as heads?

8. Mo and Flo plan to have 3 children.

   a. List the sample space using B for boy and G for girl. *Hint:* There are 8 equally likely outcomes in the sample space.
   b. What is the probability that Mo and Flo will have more boys than girls?
   c. What is the probability that Mo and Flo have exactly two girls?
   d. What is the probability that Mo and Flo have at least two girls?
   e. What is the probability that the genders of the three children alternate?
   f. What is the probability that all children are the same gender?

Answers

You Try It $5.1.A$

a. $S = \{R, B, G\}$ and $n(S) = 3$
b. $E = \{B, G\}$ and $n(E) = 2$

You Try It $5.1.B$

$S = \{RB, RW, BR, BW, WR, WB\}$ and $n(S) = 6$.

Once a color has been selected first, that same color cannot be selected second because the ball is not replaced.

You Try It $5.1.C$

$\frac{2}{8} = \frac{1}{4} = 0.25 = 25\%$

You Try It $5.1.D$

a. $\frac{2}{8} = \frac{1}{4} = 0.25 = 25\%$
b. $\frac{2}{5} = 0.4 = 40\%$

You Try It $5.1.E$

$\frac{12}{52} = \frac{3}{13} \approx 0.23 \approx 23\%$

You Try It $5.1.F$

$\frac{3}{4} = 0.75 = 75\%$

This page titled 5.1: Theoretical Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

5.2: Empirical Probability and the Law of Large Numbers

Learning Objectives

- Differentiate theoretical and empirical probability.
- Calculate empirical probability.
- Explain the *Law of Large Numbers*.

Introduction

Probability began with the idea that it is based on analyzing the equally likely outcomes in a sample space. For example, the probability that a coin lands heads up is $\frac{1}{2} = 0.5 = 50\%$ because the sample space contains two equally likely outcomes, and only one of them -- heads -- belongs to the event.

However, data from real experiments do not always match these theoretical expectations. If you flip a coin 10 times, you might observe heads 40%, 60%, or even 70% of the time. Such variation is normal and expected whenever probability is measured through actual trials. This distinction leads us to two important concepts: theoretical probability and empirical probability.

Theoretical probability describes what *should* happen in the long run. It is calculated by analyzing the structure of the sample space and imagining the experiment repeated infinitely many times. It is the ratio of the number of outcomes in the event to the number of outcomes in the sample space.

Empirical probability describes what *actually* happens in practice. It is based on observed data collected from a finite number of trials. It is the relative frequency of an event after the experiment has been performed. Because it relies on real outcomes rather than idealized ones, empirical probability is also called *experimental probability*, and the two terms may be used interchangeably.

Together, these ideas highlight the difference between mathematical prediction and real-world observation -- an essential theme in understanding probability.

Empirical Probability

Empirical probability is based on performing an actual experiment -- flipping coins, spinning spinners, drawing jellybeans from a jar, and similar activities. It represents the **relative frequency** of an event: the number of times the event occurs divided by the total number of trials. In other words, empirical probability reflects what *actually happens* when an identical experiment is repeated many times.

Example 5.2.1

Lawrence has a standard 52-card deck and wants to estimate the probability of drawing a single Queen.

The theoretical probability of drawing a Queen is $\frac{4}{52} = \frac{1}{13} \approx 0.0769$ or about 7.69%, because there are 4 Queens in a deck of 52 equally likely cards.

Lawrence performs an experiment 25 times by drawing a card at random, recording it, and replacing it. If he draws a Queen 3 times in these 25 trials, then the empirical probability is

$$\frac{3}{25} = 0.12 = 12\%.$$

This result differs from the theoretical probability. This is to be expected when working with a limited number of trials.

Example 5.2.2

A survey was conducted to determine the programs that students were taking. The results are shown in the table.



What is the probability that a student is enrolled in the Education program?

Solution

The total number of trials (students surveyed) is $3 + 8 + 2 + 12 + 10 = 35$. Of these, 10 were in the Education program. The fraction of trials that resulted in the event "*Education* " gives the empirical probability:

$$P(\text{Education}) = \frac{10}{35} = \frac{2}{7} \approx 0.2857 \text{ or } 28.57\%.$$

You Try It 5.2.A

In a given week, a veterinarian recorded how many times she treated each type of animal.

Number of Animals Treated



What is the probability that the next animal she treats is a cat?

Answer available at the end of this section.

Using the *Law of Large Numbers* to Estimate Theoretical Probability

Empirical probability is based on performing an experiment and observing how often an event occurs. By repeating the same process many times -- flipping coins, spinning spinners, drawing cards, or selecting jellybeans -- we can calculate the relative frequency of an event. This relative frequency serves as an estimate of the event's theoretical probability, the probability we would expect if the experiment could be repeated infinitely many times.

The connection between empirical results and theoretical probability is explained by the ***Law of Large Numbers***. This law states that as a probabilistic process is repeated a large number of times, the empirical probabilities (relative frequencies) of the outcomes will get closer and closer to their theoretical probabilities. In practice, this means that the more trials we perform, the more accurate our estimate becomes. Although we can never conduct infinitely many trials, a sufficiently large number of repetitions allows empirical probability to approximate theoretical probability quite well.

This principle is especially useful when theoretical probability is difficult to compute -- such as when the number of possible outcomes is extremely large -- or impossible to compute because the outcomes are not equally likely. In such cases, empirical probability provides a practical method for estimating the true probability of an event.

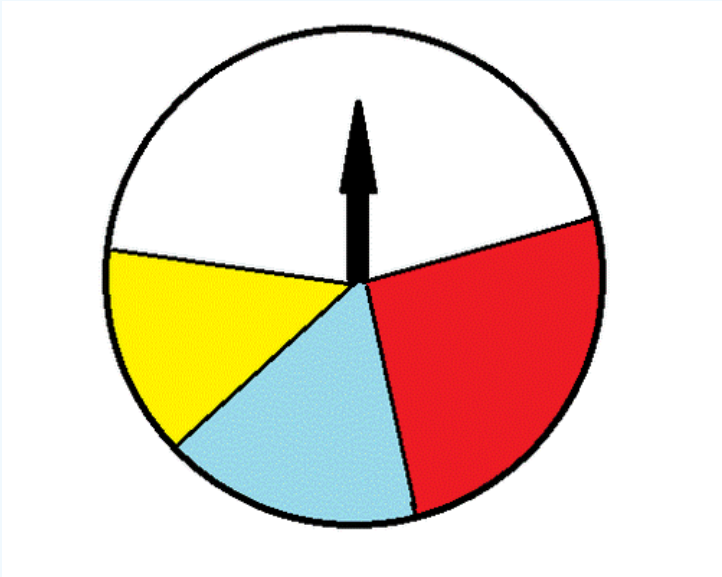
Consider the following example. Suppose we want to estimate the probability that a thumbtack lands point-up when tossed.

A thumbtack can land point-up or on its side, but these outcomes are not equally likely, so theoretical probability cannot be determined by simple counting. Instead, we can toss a thumbtack many times (or toss many thumbtacks simultaneously) and

record how often it lands point-up. The resulting relative frequency gives an empirical estimate of the theoretical probability of the event.

Example 5.2.3

A game uses a spinner like the one shown below. When spun, the arrow lands on either a white, yellow, blue, or red region. Let's describe how we could use the *Law of Large Numbers* to estimate the probability of landing on the white region.



The regions do not appear to be the same size, so the four outcomes are not equally likely. We cannot apply our theoretical probability formula here unless we are told the areas of the regions or the measures of the central angles in the circular spinner. Instead, we can spin the spinner many times and record the results:

Spinner Results



Using the data in the table, the spinner landed on "white" 18 times out of a total of $18 + 12 + 6 + 4 = 40$ times. The empirical probability that the spinner lands on white is $\frac{18}{40} = 0.45$ or 45%.

This seems a reasonable estimate for the theoretical probability as the white region makes up slightly less than half the spinner. Randomness in the results will mean that we always get an approximation of the true probability, but the more trials we do, the more confidence we can have that the experimental probability will match the theoretical probability.ere.

Example 5.2.4

Jennifer wants to compare the theoretical probability that a number cube lands on '3' with the empirical probability observed after performing 50 trials.

A fair six-sided die has 6 equally likely outcomes, and only one of them is the number '3.' Therefore, the theoretical probability is $\frac{1}{6} \approx 0.1667$ or about 17%.

To compute an empirical probability, Jennifer rolls the number cube 50 times and records each outcome.



From her data, the number cube landed on '3' in 9 out of 50 rolls. The empirical probability is $\frac{9}{50} = 0.18$ or 18%.

Although 18% does not match the theoretical probability exactly, it is reasonably close. This difference is expected when the number of trials is limited. According to the *Law of Large Numbers,* as the number of trials increases, the empirical probability tends to move closer to the theoretical probability. If Jennifer repeated the experiment 100, 1,000, or 10,000 times, the empirical probability would likely approach $\frac{1}{6}$.

We also can use empirical probability to confirm theoretical probability to see for ourselves whether something happens as often as we'd expect it to. If it doesn't, we may need to reconsider our thinking.

Example 5.2.5

Nadia and Peter are playing a game with a six-sided die, but Peter keeps winning, and Nadia begins to suspect that the die may be unfair. In particular, she notices how often Peter rolls a 6, so she decides to run an experiment. Nadia rolls the die 100 times, records each result, and organizes her data into a frequency table.



After reviewing the table, it becomes clear that something unusual may be happening: the outcome '6' appears roughly twice as often as the other numbers. This pattern suggests that the die could be weighted or otherwise biased. However, even with such a strong pattern, Nadia cannot conclude with absolute certainty that the die is unfair. Random variation can still produce surprising results, especially with a relatively small number of trials.

Because of this, Nadia must speak in terms of likelihood, not certainty. To strengthen her conclusion, she could repeat the experiment with more trials -- for example, 500 or 1,000 rolls. If the relative frequency of rolling a '6' continues to stand out as an outlier while the other outcomes remain roughly balanced, the evidence that the die is unfair becomes much stronger. This approach aligns with the *Law of Large Numbers*, which tells us that empirical probabilities tend to stabilize and reveal underlying patterns as the number of trials increases.

Example 5.2.6

Stanley picks a card from a standard deck of cards and gets heart. He returns the card to the deck, picks a second time, and gets another heart. Stanley repeats this process a total of five times and gets a heart in four of those trials. Should Stanley conclude that the probability of selecting a heart is $\frac{4}{5} = 0.80 = 80\%$?

Solution

No. Stanley's reasoning is incorrect. He should not make a conclusion about the true probability of selecting a heart based only on five trials. If Stanley wants to estimate the chance of selecting a heart using empirical probability, he should perform the experiment a lot more than five times.

You Try It 5.2.B

Juan suspects that his lucky coin is weighted so that he gets heads more often than tails in a coin toss. The reason he is suspicious is because he seems to get several heads in a row when he tosses it. He conducts an experiment and gets the following results.



Do you think he should conclude the coin is unfair? Explain.

Answer available at the end of this section.

Section Summary

Empirical probability uses observed data from an experiment to estimate how likely an event is. It is calculated as the relative frequency of the event -- how often the event occurs divided by the total number of trials.

The *Law of Large Numbers* explains the relationship between empirical and theoretical probability. As an experiment is repeated many times, the empirical probability of an event tends to move closer to the event's theoretical probability. In other words, more trials lead to more reliable estimates.

Together, these ideas show how actual experimental results can be used to approximate theoretical probabilities, especially in situations where theoretical calculations are difficult or impossible.

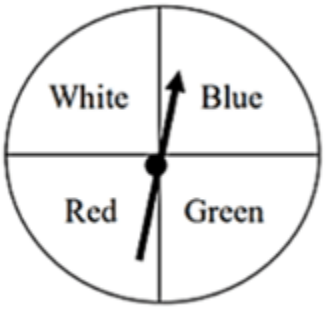
Section 5.2 Review Problems

1. According to the results of a survey, 200 people said they prefer cats, 160 said they prefer dogs, while 40 said they don't like either pet. Find the empirical probability that a person prefers dogs. Give the probability as a fraction in simplest form, as a decimal, and as a percent.

2. Find the empirical probability that the next car passing you on the highway is red if the previous 25 car colors were as follows:



3. Moya is playing a board game with the spinner shown below. The spinner has 4 sections, each the same size. Moya recorded the results of her spins in a table as she played the game.





   a. If Moya's spinner is fair, find the theoretical probability that Moya spins '*green*.'
   b. Find the empirical probability that Moya spins '*green*.'
   c. Which color has an empirical probability closest to its theoretical probability?
   d. Decide whether you think this spinner is fair. Explain your choice.

4. Determine whether you could calculate the theoretical probability of each event based on your knowledge of the possible outcomes, or whether you would have to find the empirical probability:
   a. Flipping a coin three times in a row and getting three heads.
   b. Pulling a nickel from your pocket when you know you have three nickels and five dimes in your pocket.
   c. Guessing the right answer on a four-option multiple choice question.
   d. Guessing the right answer on a free-response question.
   e. Sharing a birthday with one of your three best friends.
   f. Getting a flat tire while driving home.
   g. Having your left front tire be the one that goes flat, whenever you *do* get your next flat tire.
   h. Tossing a can of soda and it landing on its side.

Answers

You Try It $5.2.A$

$\frac{32}{90} = \frac{16}{45} \approx 0.3556$ or 35.56%.

You Try It $5.2.B$

The 40 tosses resulted in 23 heads and 17 tails. So $P(H) = \frac{23}{40} = 0.575 = 57.5\%$ and $P(T) = \frac{17}{40} = 0.425 = 42.5\%$. For the coin to be fair, we expect heads around 50% of the time. While 57.5% is higher than 50%, this difference may be due to random chance - especially since Juan performed the experiment only 40 times. Further experimentation would help to get a more certain answer to whether Juan's coin is weighted.

This page titled 5.2: Empirical Probability and the Law of Large Numbers is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

5.3: Using Sets and Venn Diagrams to Represent Events

Learning Objectives

- Represent events with Venn diagrams.
- Perform fundamental set operations.
- Determine probabilities of simultaneous events.
- Determine probabilities of alternative events.

Introduction

To this point, we have focused on finding the probability of a single event - one outcome or one collection of outcomes from a sample space. But many real situations involve two events happening at the same time or either one of two events happening. To describe these situations clearly, we need a way to organize and compare how different events overlap or combine.

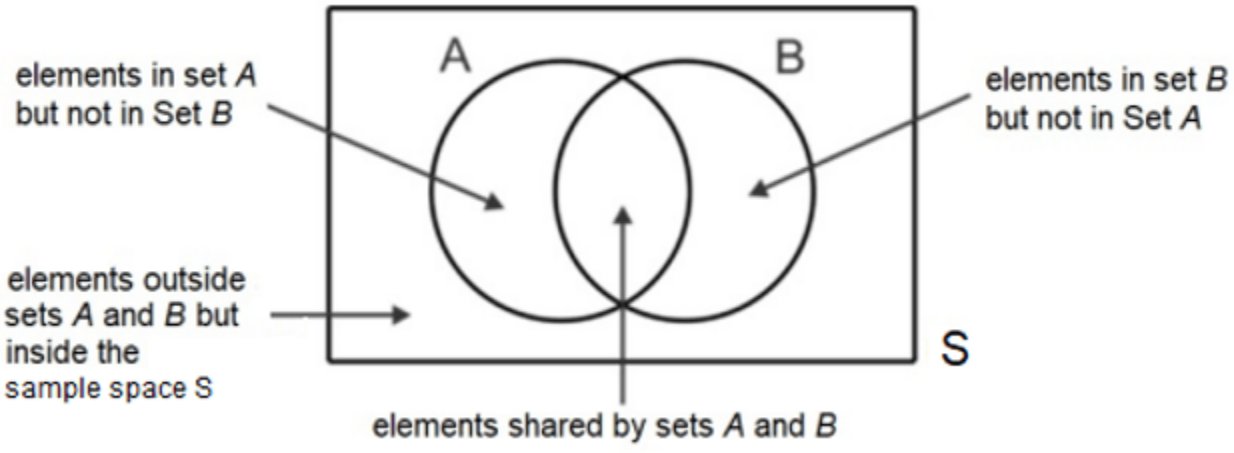
A Venn diagram is a visual tool that helps show how outcomes are arranged within two or more events. It allows you to see which outcomes belong to both events, which belong to only one event, and which belong to neither. In this section, you will learn how to represent outcomes using intersections and unions, how to read these relationships from a Venn diagram, and how to determine probabilities based on these combined events. These ideas will help you analyze more complex probability questions where events interact or overlap.

Venn Diagrams

A **Venn diagram** is a visual way to show how outcomes are related across two or more events. Each event is represented by a circle, and the rectangle around the circles represents the entire sample space of possible outcomes.

When two circles overlap, the overlap shows outcomes that belong to both events. Outcomes that fall only inside one circle belong to just that event, and outcomes outside both circles belong to neither.

In the diagram for events A and B below, you can see that both circles lie inside the sample space S. Because the circles overlap, events A and B share some outcomes. This shared region represents the *intersection* of events - the outcomes that occur in both A and B.

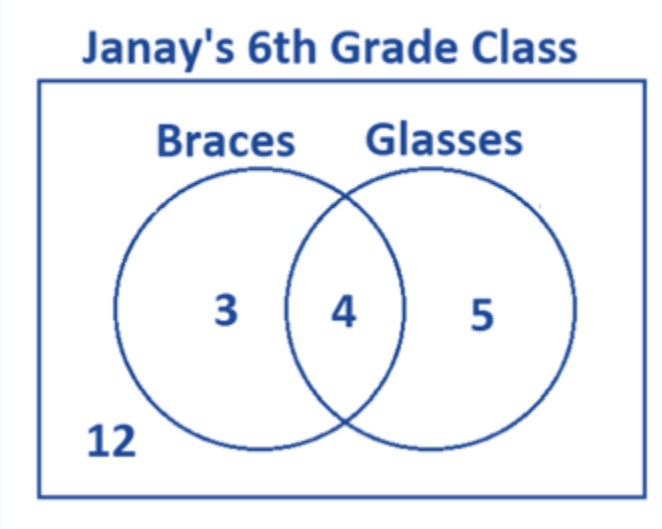


The next example uses a Venn diagram to organize outcomes from a real-world situation, making it easier to see how events can

overlap or remain separate within a single sample space.

Example 5.3.1

Janay organized the students in her 6th grade class based on whether they wear braces and whether they wear glasses. The Venn diagram shows summarizes all students in the class and counts how many fall into each category.



From the diagram, we can draw several conclusions about the class:

- Students with both: 4 students wear both braces and glasses.
- Students with braces: 3 students wear braces only, and 4 wear both, for a total of $3 + 4 = 7$.
- Braces only: 3 students wear braces but not glasses.
- Students with glasses: 5 students wear glasses only, and 4 wear both, for a total of $4 + 5 = 9$.
- Glasses only: 5 students wear glasses but not braces.
- Braces or glasses: all students in either circle add up to $3 + 4 + 5 = 12$.
- Neither braces nor glasses: 12 students fall outside both circles.
- Total students: adding all regions gives $12 + 3 + 4 + 5 = 24$ students in Janay's class.

Once the outcomes are organized in the Venn diagram, probabilities can be directly calculated from counting the outcomes in each region of the diagram and comparing them to the total number of students in the sample space $n(S) = 24$.

For example, the probability that a randomly chosen student wears braces is

$$P(\text{wears braces}) = \frac{n(\text{wears braces})}{n(S)} = \frac{7}{24}.$$

The probability that a student wears braces or glasses is

$$P(\text{wears braces or glasses}) = \frac{n(\text{wears braces or glasses})}{n(S)} = \frac{12}{24} = \frac{1}{2}.$$

Set Operations with Events

In probability, we often need to describe how events relate to one another inside a sample space. To do this, we use **set operations**, which allow us to combine events, compare them, or form new ones. These operations help us understand how outcomes can fall inside one event, inside another, in both, or in neither.

The main set operations you will use are

- **Complement:** outcomes not in an event.
- **Intersection:** outcomes that belong to *both* events at the same time.
- **Union:** outcomes that belong to *at least one* of the events.

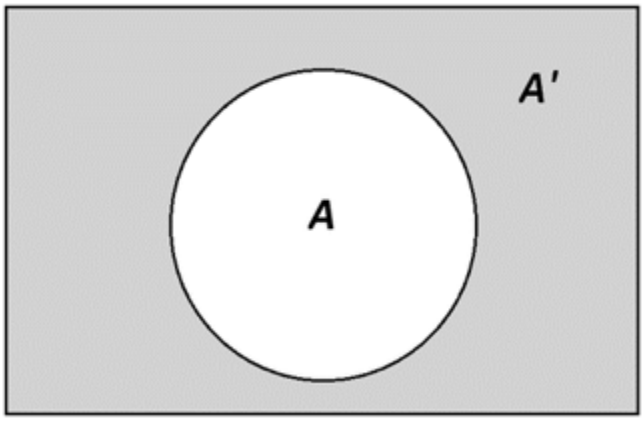
Each operation gives a different way to describe how events are connected within a sample space. You will explore each operation individually and see how they help organize outcomes and calculate probabilities.

The Complement of an Event

The **complement** of an event includes all outcomes in the sample space that are not part of event A. It can be represented in a Venn diagram by everything outside of A.

We use the symbol A' to denote the complement, often read as "not A" or "A prime."

In the Venn diagram below, the shaded region shows A': all outcomes in the sample space that do not belong to event A.



Example 5.3.2

The sample space for rolling a fair die is $S = \{1, 2, 3, 4, 5, 6\}$.

If event A is "rolling an odd number," then $A = \{1, 3, 5\}$, and there are 3 outcomes in A.

The complement of A consists of all outcomes in the sample space that are not odd.

Therefore, $A' = \{2, 4, 6\}$, and this set also contains 3 outcomes.

You Try It 5.3.A

For an experiment, you randomly select a letter from the letters in the word **JONES**.

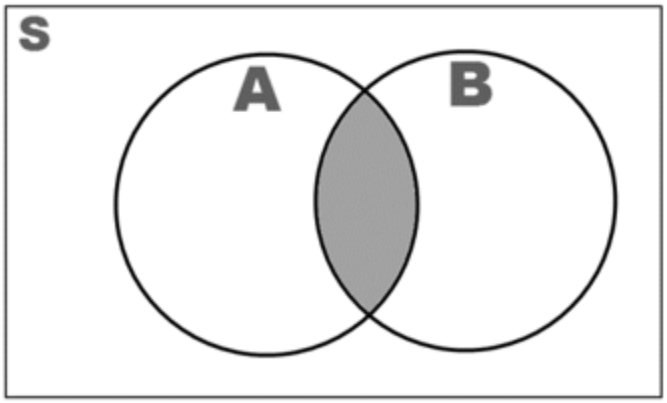
If event A is "selecting a vowel," find A'.

Answer available at the end of this section.

Intersection of Events

Sometimes we form a new event by looking at where two existing events overlap. Suppose a sample space contains events A and B, and some outcomes belong to both. In this situation, the events are **not mutually exclusive** because they share at least one outcome.

The **intersection** of events $A \cap B$ is the set of all outcomes that belong to both A and B. In a Venn diagram, this is the region where the two circles overlap. The symbol $\cap$ represents the intersection, and it is often associated with the word "**and.**" The shaded region in the diagram below shows $A \cap B$: every outcome that belongs to both events at the same time. To determine how many outcomes are in the intersection, you simply count the outcomes that appear in both sets.



Example 5.3.3

Suppose set $A = \{c, a, t\}$ and set $B = \{a, r, t\}$.

Find $A \cap B$.

Solution

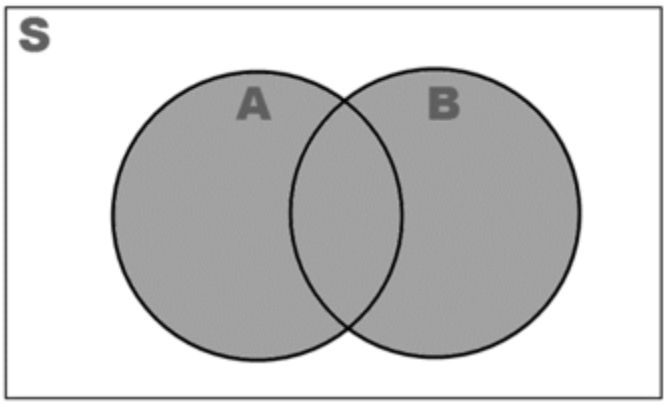
The intersection $A \cap B$ contains all letters that appear in both words. Both sets share the letters a and t.

So $A \cap B = \{a, t\}$.

Union of Events

The **union** of two events A and B is a new event that includes every outcome that is in A, in B, or in both. The symbol $\cup$ represents the union, and it is commonly associated with the word **"or."**

In the Venn diagram shown below, the union is represented by shading both circles, including their overlap. This shaded region represents all outcomes that belong to at least one of the two events.



When finding the union, it is important to avoid listing and counting shared outcomes twice. The overlap -- the intersection -- must be counted only once when determining the total number of outcomes in $A \cup B$.

Example 5.3.4

Suppose set $A = \{c, a, t\}$ and set $B = \{a, r, t\}$.

Find $A \cup B$.

Solution

Theunion $A \cup B$ contains all letters that appear in at least one of the words That is, all the letters that appear in A or B or both A and B.

So $A \cup B = \{c, a, t, r\}$.

You Try It 5.3.B

Suppose the sample set set is the letters in the word ***elastic***:

$$S = \{\ \mathrm{e},\ \mathrm{l},\ \mathrm{a},\ \mathrm{s},\ \mathrm{t},\ \mathrm{i},\ \mathrm{c}\ \}.$$

Consider these sets: $A = \{\ \mathrm{s},\ \mathrm{c},\ \mathrm{a},\ \mathrm{l},\ \mathrm{e}\ \}$ $B = \{\ \mathrm{c},\ \mathrm{a},\ \mathrm{t}\ \}$

a. Find $A \cup B$.
b. Find $A \cap B$.
c. Find A'.

Answer available at the end of this section.

Example 5.3.5

Consider the experiment of rolling a single fair die.

- Event A: observe an even number.
- Event B: observe a number less than or equal to 3.

a. List the outcomes in the sample space S, event A, and event B.

Solution

$S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$, and $B = \{1, 2, 3\}$

b. List the outcomes in A' and find $n(A')$.

Solution

The complement A' contains all outcomes in the sample space that are not even:

$A' = \{1, 3, 5\}$ and $n(A') = 3$.

c. List the outcomes in $A \cup B$ and find $n(A \cup B)$.

Solution

An outcome is in the union $A \cup B$ if it belongs to event A, event B, or both.

$A \cup B = \{1, 2, 3, 4, 6\}$ and $n(A \cup B) = 5$.

d. List the outcomes in $A \cap B$ and find $n(A \cap B)$.

Solution

An outcome is in the intersection $A \cap B$ if it belongs to both events.

$A \cap B = \{2\}$ and $n(A \cap B) = 1$.

e. Calculate $P(A')$, $P(A \cup B)$, and $P(A \cap B)$.

Solution

Probability is the number of outcomes in the event divided by the number of outcomes in the sample space.

$P(A') = \frac{3}{6} = \frac{1}{2}$

$P(A \cup B) = \frac{5}{6}$

$P(A \cap B) = \frac{1}{6}$.

Example 5.3.6

Suppose we add a new set to the sets used in Example 5.3.5:

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$A = \{2, 4, 6\}$$

$$B = \{1, 2, 3\}$$

$$C = \{6\}$$

a. List the outcomes in $(A \cap C)'$.

Solution

Simplify inside the parentheses first. An outcome is in the intersection $A \cap C$ if it belongs to both events.

$A \cap C = \{2, 4, 6\} \cap \{6\} = \{6\}$

Now, find the complement of that result.

$(A \cap C)' = \{6\}' = \{1, 2, 3, 4, 5\}$

b. List the outcomes in $A \cup B \cup C$.

Solution

We need to find the union of the three sets. To do so, it is a good idea to use the *associative property* by first finding the union of events A and B, and then finding the union of that result with event C.

$$\begin{aligned}(A \cup B) \cup C &= (\{2, 4, 6\} \cup \{1, 2, 3\}) \cup \{6\} \\ &= \{1, 2, 3, 4, 6\} \cup \{6\} \\ &= \{1, 2, 3, 4, 6\}\end{aligned}$$

c. List the outcomes in $A \cap B \cap C$.

Solution

We need to find the intersection of the three sets. Again, use the *associative property* by first finding the intersection of events A and B, and then intersecting the result with event C:

$$\begin{aligned}(A \cap B) \cap C &= (\{2,4,6\} \cap \{1,2,3\}) \cap \{6\} \\ &= \{2\} \cap \{6\} \\ &= \{\quad\} \\ &= \emptyset\end{aligned}$$

There are no outcomes in $A \cap B \cap C$ so the result is the empty set.

Table of Set Operations

Here is a table of operations that can be used to describe each region of the Venn diagram with two events that are not mutually exclusive.

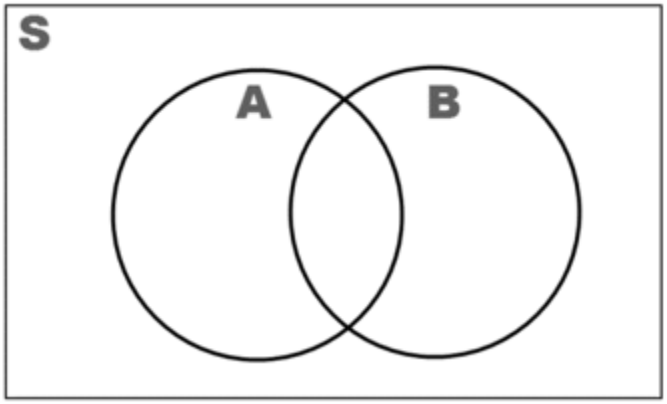


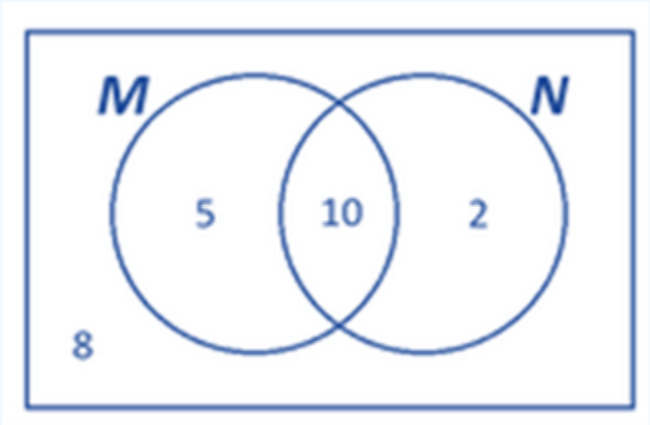
Table of Set Operations

Symbol	Description
A	outcomes in event A
$A \cap B$	outcomes in the *intersection* of events A and B (all outcomes that are in both events)
$A \cap B$	outcomes in the *union* of events A and B (all outcomes that are in one or both sets)
A'	outcomes in the *complement* of event A (all outcomes of the sample space not in event A)
$(A \cup B)'$	outcomes in the *complement of the union* of events A and B (all outcomes of the sample space that are neither in event A nor event B)
$(A \cap B)'$	outcomes in the *complement of the intersection* of events A and B (all outcomes of the sample space that are not in both A and B

Symbol	Description
$A \cap B'$	outcomes in the *intersection* of event A and the *complement* of event B (all outcomes of the sample space that are in A but not in B)

Example 5.3.7

The Venn diagram shows the number of outcomes in events A and \(B. Use the Venn diagram to find each indicated probability. *A*ssume that all outcomes are equally likely.



a. $P(M)$

Solution

The sample space contains $5 + 10 + 2 + 8 = 25$ outcomes. There are $5 + 10 = 15$ outcomes in event M so $P(M) = \frac{15}{25} = \frac{3}{5}$.

b. $P(N')$

Solution

There are $8 + 5 = 13$ outcomes *not* in event N, so $P(N') = \frac{13}{25}$.

c. $P(M \cap N)$

Solution

There are 10 outcomes in *both* event M and N so $P(M \cap N) = \frac{10}{25} = \frac{2}{5}$.

d. $P(M \cup N)'$

Solution

There are 8 outcomes in neither event M and N so $P(M \cup N)' = \frac{8}{25}$

e. $P(M \cap N')$

Solution

There are 5 outcomes in event M but not in event N so $P(M \cap N') = \frac{5}{25} = \frac{1}{5}$.

Problem Solving with Venn Diagrams

Venn diagrams are a powerful visual tool for organizing information outcomes in probability. They help you see how events relate to one another within a sample space -- where they overlap, where they differ, and where outcomes fall outside both events. They allow you can clearly identify which outcomes belong to each event and how many outcomes each region contains. This visual structure makes it easier to compute probabilities, avoid double-counting, and understand how events interact.

Example 5.3.8

At a school of 500 students, there are 125 students in the drama club, 257 students who play sports, and 52 students who are in the drama club and play sports.

Create a Venn diagram to illustrate this information.

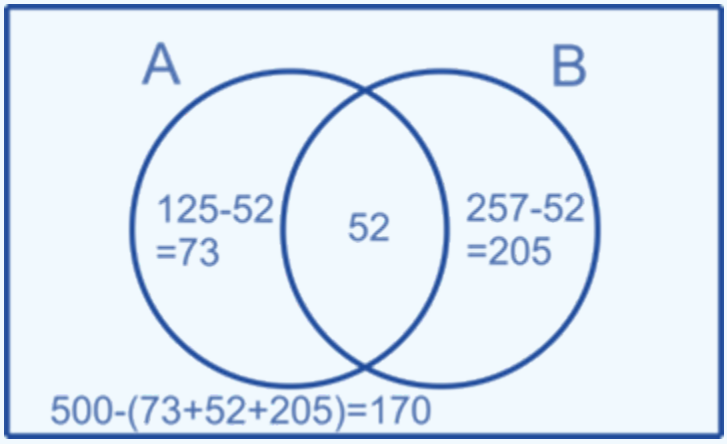
Solution

Let Event A be the students in the drama club and Event B be the students who play sports.

Because 52 students are in *both* groups, the most efficient approach is to place this overlap of 52 in the center of the diagram first.

Then, subtract this shared group from each event total to find how many students belong only to drama $(125 - 52 = 73)$ or only to sports $(257 - 52 = 205)$.

Finally, subtract all students represented in the circles from the total school population to determine how many students are outside both events: $500 - (52 + 73 + 205) = 170$.



Example 5.3.9

Suppose a student from the high school in Example 5.3.8 is selected at random. Find the probability of these events.

a. Find the probability of selecting a student in sports but not in drama club.

Solution

There are 205 students in sports but are not in drama club.

$$\frac{205}{500} = \frac{41}{100} = 0.41 = 41\%$$

b. Find the probability of selecting a student who is in the drama club or in sports.

Solution

There are $73 + 52 + 205 = 330$ students who are in drama club or in sports.

$$\frac{330}{500} = \frac{33}{50} = 0.66 = 66\%$$

c. Find the probability of selecting a student who is in exactly one of drama club or sports.

Solution

There are 73 students in drama club but not sports and 205 students in sports but not drama club. This is a total of $73 + 205 = 278$ students who are in exactly one of dram club or sports.

$$\frac{278}{500} = \frac{139}{250} = 0.556 = 55.6\%$$

d. Given the student is in the drama club, what is the probability of selecting a student who plays sport?

Solution

The sample space is restricted to those 125 students in drama club. Of these, only 52 are also in sports.

$$\frac{52}{125} = 0.416 = 41.6\%$$

Example 5.3.10

In a survey of 150 students, it was found that

- 80 students have laptops,
- 110 students have cell phones,
- 125 students have iPods,
- 62 students have both a laptop and a cell phone,
- 58 students have both a laptop and iPod,
- 98 students have both a cell phone and an iPod, and
- 50 students have all three items.

Draw a Venn diagram to illustrate the relationship between students who have laptops, cell phones, and iPods.

Solution

There are three sets of students: laptops, cell phones, and iPods. A Venn diagram for three sets results in 8 regions. We will fill the 7 regions inside the three circles, then find how many students are outside all circles.

1. Start with the overlap of all three circles. These are 50 students with all three items (laptop, cell phone, iPods.)

   So, the "all three items" region is 50.

2. Fill the “exactly two items” regions. Use the given pairwise totals and subtract the 50 who have all three:

   - Laptop and cell phone only $= 62 - 50 = 12$
   - Laptop and iPod only $= 58 - 50 = 8$
   - Cell phone and iPod only $= 98 - 50 = 48$

So the “exactly two items” regions are 12, 8, and 48.

3. Now use the single-item totals and subtract all overlapping regions in that circle:

   - Laptop only: $80 - (12 + 8 + 50) = 80 - 70 = 10$
   - Cell phone only: $110 - (12 + 48 + 50) = 110 - 110 = 0$
   - iPod only: $125 - (8 + 48 + 50) = 125 - 106 = 19$

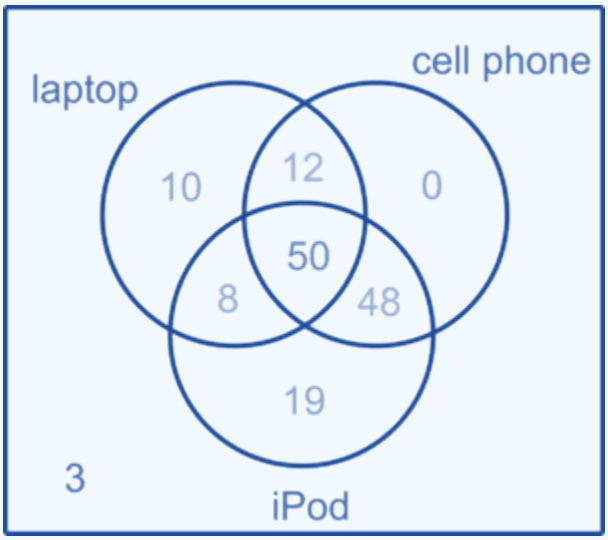
   So the “only one item” regions are 10, 0, and 19.

4. Find how many have none of the items by adding all students inside the circles and subtracting from the total number of students surveyed

   There are $10 + 0 + 19 + 12 + 8 + 48 + 50 = 147$ students who have at least one of these items.

   This leaves $150 - 147 = 3$ students in the "none of the items."

5. Place the number of students in each region of the Venn diagram as shown:



   - Laptop only: 10
   - Cell phone only: 0
   - iPod only: 19
   - Laptop & cell only: 12
   - Laptop & iPod only: 8
   - Cell & iPod only: 48
   - All three: 50
   - None: 3

Section Summary

Venn diagrams provide a visual way to represent these relationships. By shading regions that correspond to complements, unions, and intersections, Venn diagrams help organize information, prevent double-counting, and make it easier to identify which outcomes belong to each event. They are an effective tool for solving probability problems and understanding how events overlap within a sample space.

The complement of event A, written A', consists of all outcomes in the sample space that are not in event A.

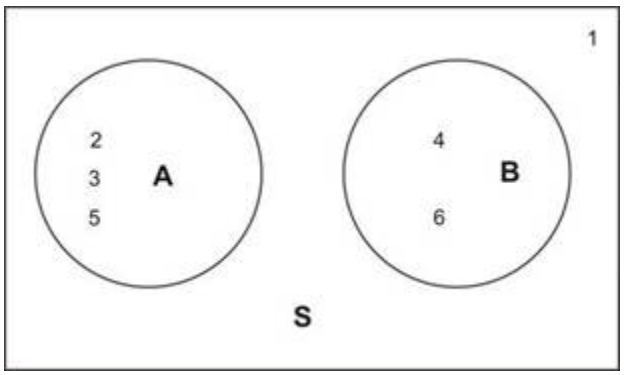
The intersection of two events $A \cup B$ includes only the outcomes that are in both event A and event B. An intersection represents

an 'and' relationship.

The union of two events $A \cup B$ includes every outcome that is in event A, event B, or both. A union represents an 'or' relationship.

Section 5.3 Review Questions

1. The Venn diagram shows an experiment with six equally likely outcomes: $\{1, 2, 3, 4, 5, 6\}$. Events A and B are shown by the Venn diagram. Find the probabilities of each event.



   a. Find $P(A)$.
   b. Find $P(B)$.
   c. Find $P(A \cup B)'$.

2. Data from a survey of 140 students showed 37 study music, 103 play a sport, and 22 do both activities.
   a. Represent these data with a well-labeled Venn diagram.
   b. How many students study music but do not play a sport?
   c. How many students study music or play a sport?
   d. What is the *probability* of selecting student who does neither activity?

3. In a fifth-grade class of 25 students, 5 students are left-handed. Seven of the students wear glasses. Furthermore, 3 of the students are left-handed and wear glasses.
   a. Draw a complete Venn diagram to represent this situation.
   b. Find the *probability* that a student is left-handed.
   c. Find the *probability* that a student does not wear glasses.
   d. What is the *probability* that a student wears glasses but is not left-handed?
   e. What is the *probability* that a student is left-handed or wears glasses?
   f. What is the probability that student is neither left-handed nor wears glasses?
   g. Given a student is left-handed, what is the probability that a student wears glasses?
   h. Given a student doesn't wear glasses, what is the probability that a student is left-handed?

4. Consider $S = \{a, b, c, d, e, f\}$ to be the sample space of an experiment. Also, event $X = \{X = \{c, d, e, f\}$, event $Y = \{a, c, d\}$, and even $Z = \{b, e\}$. List the outcomes in each of these new events:
   a. X'
   b. $X \cap Y$
   c. $X \cup Y$
   d. $Y \cap Z$
   e. $(Y \cup Z)'$
   f. $X \cup Y \ \cup Z$

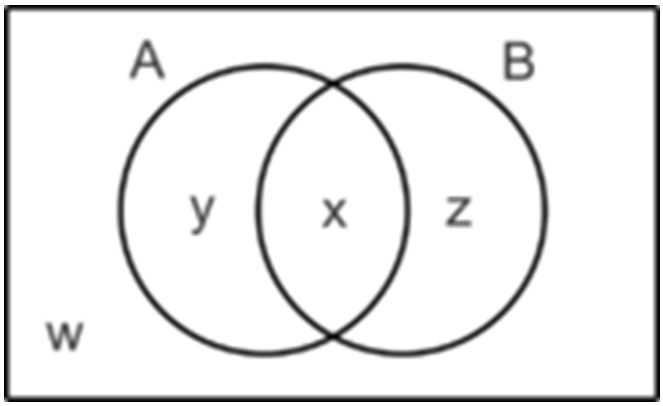
5. The sample space for an experiment is $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ with events

$$A = \{1, 2, 3, 4, 5, 6\}\ B = \{1, 2, 3, 5, 7\}\ C = \{2, 4\}\ D = \{8, 9\}$$

   List the outcomes in each compound event:

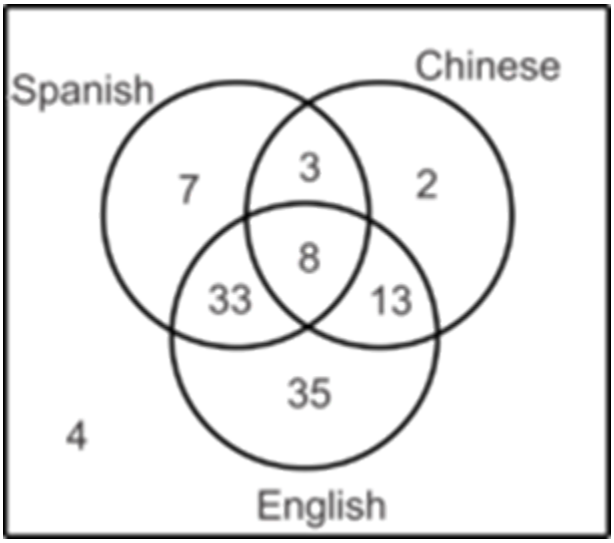
a. $A \cap B$
b. $A \cup B$
c. A'
d. $A \cap C$
e. $A \cup C$
f. $B \cap D$
g. $B \cup D$
h. B'
i. $A \cap B \cap C$
j. $(A \cup B)'$

6. Use one or more letters in the Venn diagram $\{w, x, y, x\}$ to identify the region corresponding to each of the sets or set operations.



a. A
b. $A \cup B$
c. $A \cap B'$
d. $(A \cap B)'$
e. $(A \cup B)'$
f. A'
g. $(B' \cup A$

7. The International Club at a school has 105 members, many of whom speak multiple languages. The most common languages in the club are English, Spanish, and Chinese.



Use the Venn Diagram to determine the *probability* of selecting a student who

a. does not speak English.
b. speaks Spanish and English but not Chinese
c. speaks Spanish given that he/she speaks English.
d. speaks English given that he/she speaks Chinese.

8. A veterinarian surveys 25 of her patrons. She discovers that 14 have dogs, 10 have cats, and 5 have fish. Four have dogs and cats, 3 have dogs and fish, and 2 have cats and fish. Only 1 person has all three kinds of pets.
   a. Draw a Venn diagram to organize the data.
   b. How many patrons have none of these pets?
   c. What is the probability that a patron has a dog and fish but not a cat?
   d. What is the probability that a patron has exactly one kind of these pets?
   e. Given that a patron owns a cat, what is the probability of owning a fish?

9. In a class of 300 seniors, 125 study Spanish, 95 study Chemistry, 165 study Mathematics, 18 study Spanish and Chemistry, 75 study Chemistry and Math, 20 study Math and Spanish and 15 study all three subjects.
   a. Make a Venn diagram to illustrate these data.
   b. Find the *probability* that a student selected at random.
      i. studies only Spanish.
      ii. studies Math and Chemistry but not Spanish.
      iii. studies none of these subjects.
      iv. studies exactly one of these

Answers

You Try It $5.3.A$

$A' = \{J, N, S\}$

You Try It $5.3.B$

a. $\{s, c, a, l, e, t\}$
b. $\{c, a\}$
c. $\{t, i\}$

This page titled 5.3: Using Sets and Venn Diagrams to Represent Events is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by Andy Jones.

- 13.2: Union, Intersection, and Complement by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

5.4: Contingency Tables, Conditional Probability, and Independence

Learning Objectives

- Compute probabilities from contingency tables.
- Find and interpret conditional probability.
- Assess independence using conditional probabilities.

Introduction

A contingency table is a structured way to display data for two categorical variables at the same time. Contingency tables are especially useful in probability because they allow you to quickly identify totals for each category, compare groups, and compute probabilities directly from the counts in the table. They also provide a clear framework for checking whether two events are related using these probabilities. By organizing information visually and numerically, contingency tables make it easier to analyze relationships between events and draw meaningful conclusions.

Reading a Contingency Table

When a survey collects two pieces of categorical information from each person, the results can be organized in a **contingency table**, or **two-way table**. One variable is organized by rows, and the other by columns, creating a grid that shows how often each combination of categories occurs. The table shows how many people fall into each combination of responses.



The table above displays whether people have cable TV and whether they took a vacation in the past year. Each cell tells you how many people fit a specific pair of responses—for instance, 97 people have cable TV and took a vacation, while 17 people do neither of these.

Row and column totals summarize how many people fall into each single category: 135 people have cable TV, 31 do not, and 166 people were surveyed in total. Likewise, 111 people took a vacation and 55 did not.

A contingency table makes it easy to define events and compute probabilities directly from the counts in the table. By reading across rows, down columns, or within individual cells, you can quickly determine how likely different combinations of outcomes are.

Example 5.4.1

Use the table above to find the probability that a randomly selected person took a vacation last year.

Solution

Let A be the event that a person took a vacation. To find $P(A)$, we use the total number of people surveyed as the sample space. Since 166 people participated in the survey, the sample space contains 166 outcomes.

From the contingency table, the total number of people who took a vacation is 111. Therefore,

$$P(A) = \frac{n(A)}{n(S)} = \frac{111}{166} \approx 0.6687$$

So the probability that a randomly chosen person took a vacation last year is approximately 66.87%.

Example 5.4.2

Find the probability that a randomly selected person took a vacation last year *and* has cable TV.

Solution

Let $A =$ the event that a person took a vacation, and $B =$ the event that a person has cable TV.

We want the probability of both events occurring, written $P(A \cap B)$.

From the contingency table, 97 people both took a vacation and have cable TV. Since 166 people were surveyed, the probability is

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{97}{166} \approx 0.5843$$

Thus, the probability that a randomly chosen person both took a vacation and has cable TV is about 58.43%.

Example 5.4.3

Find the probability that a randomly selected person took a vacation last year *or* has cable TV.

Solution

Again, let $A =$ the event that a person took a vacation, and $B =$ the event that a person has cable TV.

We want the probability of the union of the two events, written $P(A \cup B)$.

From the contingency table, 149 people did one, the other, or both of these things. Since 166 people were surveyed, the probability is

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{149}{166} \approx 0.8976$$

Thus, the probability that a randomly chosen person took a vacation last year or has cable TV is about 89.76%.

Example 5.4.4

Given a randomly selected person from this group has cable TV, what is the probability they took a vacation last year?

Solution

Once again, let $A =$ the event that a person took a vacation and $B =$ the event that a person has cable TV.

Since you know that the person has cable TV (B), the sample space has been restricted to the 135 people with cable TV and $n(S) = 135$.

Of these, 97 people took a vacation so $n(A \cap B) = 97$.

$$P(A \text{ given } B) = P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{97}{135} \approx 0.7185$$

Given a randomly selected person from this group has cable TV, the probability they took a vacation last year is about 71.85%.

The sample space was restricted to the people who have cable TV -- and the remainder of the sample space was ignored. This type of probability is called a *conditional probability*. We will see similar ideas in subsequent examples and more formally define it there.

Example 5.4.5

Nate recorded the topping and type of crust for all pizzas sold yesterday.



Use the contingency table to find the following probabilities.

a. If Nate sells a pizza, what is the probability it has sausage topping?

Solution

Let G be the event that a pizza has sausage topping. The sample space consists of $n(S) = 80$ pizzas. Nate sold $n(G) = 31$ sausage pizzas, so

$$P(G) = \frac{n(G)}{n(S)} = \frac{31}{80} = 0.3875$$

The probability is 38.75%.

b. If Nate sells a pizza, what is the probability it has sausage topping or thick crust?

Solution

We want the probability of the union of sausage topping with thick crust, $P(G \cup T)$.

Nate sold 31 sausage pizzas, 10 thick-crust pepperoni pizzas, and 5 thick-crust veggie pizzas. Adding these gives $n(G \cup T) = 46$ pizzas, making sure not to double-count any pizzas already included.

$$P(G \cup T) = \frac{n(G \cup T)}{n(S)} = \frac{46}{80} = 0.575$$

The probability is 57.5%.

c. If Nate sells a pizza, what is the probability it has sausage topping and thick crust?

Solution

We want the probability of the intersection of sausage topping with thick crust, $P(G \cap T)$.

Nate sold $n(G \cap T) = 15$ pizzas that were both sausage and thick crust

$$P(G \cap T) = \frac{n(G \cap T)}{n(S)} = \frac{15}{80} = 0.1875$$

The probability is 18.75%.

d. Given Nate sells a sausage pizza, what is the probability it has thick crust?

This is a conditional probability. The sample space is restricted to the 31 sausage pizzas that Nate sold: $n(S) = 31$.

Of these sausage pizzas, 15 have thick crust so $n(G \cap T) = 15$.

$$P(T \text{ given } G) = P(T|G) = \frac{n(T \cap G)}{n(S)} = \frac{15}{31} \approx 0.4839$$

The probability is about 48.39%.

You Try It 5.4.A

The table below shows the number of credit cards owned by a group of individuals. One person will be selected at random. Find the probability of each event below.



a. the person is female
b. the person is male and has exactly one credit card
c. the person is male or has no credit cards.
d. the person is female given that they have 2 or more credit cards

Answer available at the end of this section.

Conditional Probability

Conditional probability describes the likelihood of an event once additional information -- a condition -- is known. When new information restricts the set of possible outcomes in the sample space, the probability is calculated within this smaller, more specific group. For example, the probability that a student is in the honor society changes when we know the student earned all A's this semester. In that case, the sample space is limited to students with straight-A grades rather than the entire student body.

In general, conditional probability focuses on how probabilities shift when the context changes. By narrowing the sample space to only those outcomes that satisfy a given condition, we can determine how likely another event is within that restricted group. The next example illustrates how this idea works in practice.

Example 5.4.6

Consider a high school with 400 students. The contingency table shows the number of students in each grade who earned each midterm letter grade.



Suppose one student is selected at random. Find the probability that a student

a. earned an A

b. earned an A, given the student is a 9th grader.

Solution

a. The total number of students who earned an A is $20 + 18 + 24 + 18 = 80$ out of 400 students:

$$P(\text{earned an A}) = \frac{80}{400} = 0.20 = 20\%.$$

b. The sample space is restricted to 9th graders only. There are 100 9th graders and 20 of them earned an A.

$$P(\text{earned an A given 9th grader}) = \frac{20}{100} = 0.20 = 20\%.$$

The probability in part b of Example 5.4.6 is conditional because the additional information -- knowing the student is a 9th grader -- restricts the sample space. A conditional probability describes the likelihood of event B occurring *given* that event A has already occurred. The notation $P(B|A)$ is read as "*the probability of* B given A."

In this case, we calculated $P(A \mid \text{9th grader})$. To find this probability, we focused only on the 100 students who are 9th graders and then determined how many of them earned an A (20):

$$P(\text{earned an A given 9th grader}) = \frac{20}{100} = 0.20 = 20\%$$

This result can also be obtained using a formula for conditional probability:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Here, $A =$ "student is a 9th grader" and $B =$ "student earned an A."

Since $P(A \cap B) = \frac{20}{400}$ and $P(A) = \frac{100}{400}$, we have

$$P(\text{earned an A given 9th grader}) = \frac{\frac{20}{400}}{\frac{100}{400}} = \frac{20}{100} = 0.20$$

Conditional Probability

In general, the conditional probability of event B given A can be computed using the formula

$$P(B|A) = \frac{n(A \cap B)}{n(A)} = \frac{P(A \cap B)}{P(A)}.$$

Independence

Two events are **independent** when knowing that one event has occurred does not change the probability of the other event. In other words, the occurrence of event A has no effect on the likelihood of event B. We express this idea using the notation

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

which states that the conditional probability of B given A is the same as the original probability of B.

In the previous example, the probability that a randomly selected student from the entire school earned an A was

$$P(\text{earned an A}) = 0.20.$$

We also found that the probability a 9th grader earned an A was

$$P(\text{earned an A} \mid \text{9th grader}) = 0.20.$$

Because these two probabilities are equal, knowing that a student is in 9th grade does not affect the chance of earning an A. Therefore, the events “9th grader” and “earned an A” are independent.

Independence is an important concept because it tells us when two events do not influence each other. When events are independent, their relationship is simpler to analyze, and probabilities involving both events can be computed using straightforward rules.

Independent Events

Two events A and B are independent if and only if

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

Example 5.4.7

A group of people were surveyed about the type of movies they prefer. Suppose one person is selected at random from this group. The table shows the number of males and females who prefer each movie type.



You may find it useful to find the totals for each row and column. There are 25 females and 39 males in the group. The distribution of movies is 20 romantic comedies, 35 action, and 9 horror. There are 64 total people in the group.

a. Let F be the event that the person is female. Find $P(F)$.

Solution

There are 25 females out of 64 people, so $P(F) = \frac{25}{64} \approx 0.3906$.

b. Let R be the event that the person prefers romantic comedies. Find $P(F|R)$.

Solution

The probability that the person is female *given that* the person prefers romantic comedies is a conditional probability.

Among the 20 people who prefer romantic comedies, 12 are female, so $P(F|R) = \frac{12}{20} = 0.60$.

c. Are events F and R independent? Justify your answer.

Solution

To determine independence, compare $P(F)$ and $P(F|R)$. If the events are independent, these probabilities will be equal.

$P(F) \approx 0.3906$ and $P(F|R) = 0.60$. Because $P(F) \neq P(F|R)$, knowing that a person prefers romantic comedies changes the probability that the person is female. Therefore, the events “female” and “prefers romantic comedies” are not independent.

It should be noted that there is a second way to check whether F and R are independent. Compare $P(R)$ and $P(R|F)$.

From the table,

$$P(R) = \frac{20}{64} = 0.3125$$

$$P(R|F) = \frac{12}{25} = 0.48$$

Because $P(R) \neq P(R|F)$, the probability of preferring romantic comedies changes when we know the person is female. In other words, the chance of selecting someone who prefers romantic comedies is higher when we restrict the sample space to females.

Therefore, the events female and prefers romantic comedies are not independent.

Example 5.4.8

Recall the contingency table from Example 5.4.5 where Nate recorded the topping and type of crust for all pizzas sold yesterday.



Are "selling veggie" and "selling thin crust" independent?

Solution

To determine independence, compare $P(V)$ and $P(V|T')$. If the events are independent, these probabilities will be equal.

$$P(V) = \frac{15}{80} = 0.1875 \text{ and } P(V|T') = \frac{10}{50} = 0.20$$

Because $P(V) \neq P(V|T')$, knowing that a pizza has thin crust changes the probability that the pizza is veggie. Therefore, the events “selling veggie” and “selling think crust” are not independent.

Exercise 5.4.B

Recall the table from *You Try It A* about credit cards and gender.



Are the events "two or more credit cards" and "male" independent? Explain why or why not with two probabilities.

Answer available at the end of this section.

Example 5.4.9

A medicine label reports the following results from clinical trials:

- The probability of experiencing insomnia is $P(I) = 0.14$.
- The probability of experiencing a headache is $P(H) = 0.26$.
- The probability of experiencing both side effects is $P(I \cap H) = 0.05$

Are experiencing insomnia and experiencing headaches independent events?

Solution

To check for independence, compare $P(I)$ with the conditional probability $P(I|H)$. If the events are independent, these probabilities will be equal.

We are told that $P(I) = 0.14$.

$$P(I|H) = \frac{P(I \cap H)}{P(H)} = \frac{0.05}{0.26} \approx 0.1923.$$

Because $P(I) \neq P(I|H)$, the probability of experiencing insomnia changes when we know the patient has a headache. A patient who experiences headaches is more likely to experience insomnia. Therefore, the events insomnia and headache are not independent.

Section Summary

A contingency table displays the joint distribution of two categorical variables. It shows how often two events occur together and allows us to compute probabilities such as individual event probabilities, joint probabilities, and conditional probabilities.

A conditional probability is a probability computed under additional information that restricts the sample space. Instead of considering all possible outcomes, we focus only on those outcomes for which a given condition is true. Conditional probabilities are computed using

$$P(B|A) = \frac{n(A \cap B)}{n(A)} = \frac{P(A \cap B)}{P(A)}.$$

This formula reflects the idea that once event A has occurred, only the outcomes within A remain possible.

Two events are independent when knowing that one event has occurred does not change the probability of the other. Formally, events A and B are independent if and only if

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

When this equality holds, the occurrence of A has no influence on the likelihood of B. Independence is an important concept because it simplifies probability calculations and helps identify when two events do not affect each other.

Together, contingency tables, conditional probability, and independence form a foundational toolkit for analyzing relationships between categorical variables and understanding how information changes probability assessments.

Section 5.4 Review Questions

1. The contingency table shows the results of a survey of high school students. The grade and preferred pet (cat or dog) were recorded for each respondent. Suppose a person is chosen at random from this group. Let C be the event that the student prefers cats. Let T be the event that the student is in 10th grade.



   a. Find $P(C)$.
   b. Find $P(T)$.
   c. Find $P(C \cap T)$.
   d. Find $P(C \cup T)$.
   e. Find $P(C|T)$.
   f. Find $P(T|C)$.

g. Are C and T independent? Justify your answer.

2. The middle school students at your school were surveyed and classified according to grade level and response to the question "*How do you usually get to school?*" The data are summarized in the two-way table below. A student is chosen random.



a. What is the probability that the student is in 6th grade and takes the bus?
b. What is the probability that a 6th grade student takes the bus?
c. What is the probability that a student who rides the bus is in 6th grade?
d. The previous three questions each have to do with 6th grade and taking the bus. Why are the answers to these questions different?
e. Are the events "being in 6th grade" and "taking the bus" independent? Justify your answer.

3. The table shows a random sample of cyclists and the routes they prefer. Suppose a cyclist is selected randomly. Find each probability.



a. prefers the hilly path
b. is female
c. prefers the hilly path or female
d. prefers the hilly path and female
e. prefers the hilly path given the cyclist is female
f. is a female given the cyclist prefers the hilly path
g. Are "*prefers hilly path*" and "*female*" independent?

4. The fifth-grade math students at Hill Elementary recently were part of an experiment with two books. Of the 200 students, some students used Book A and the rest used Book B. At the end of the experiment, the students were given a test to measure their grade level. The results are in the table below. Study the results and use them to answer the questions.



a. Find P(a student scored below grade level).
b. Find P(a student used Book A and scored below grade level).
c. Find P(a student used Book A or scored below grade level).
d. Find P(a student scored below grade level given they used Book A).
e. Find P(a student scored below grade level given they used Book B).
f. Does it appear that '*scoring below grade level*' and '*using Book A'* are independent? Explain clearly with probabilities.

5. A survey of 350 students at a university revealed the following data about class standing and place of residence. One student is selected at random.



a. What is the probability the student lives with their parents?
b. What is the probability the student is a freshman and lives in an apartment?
c. What is the probability the student is a junior or lives in in the dormitory?
d. What is the probability that a student is a sophomore if the student lives in an apartment?
e. What is the probability that a student lives with parents if the student is a freshman?

6. For a group of people, the probability of having blond hair is 25%. The probability of having blond hair and blue eyes is 10%. What is the probability that person has blue eyes given they have blond hair?

7. At a college, 40% of students belong to a club and 50% work part time. Five percent of the students work part time and belong to a club. Are being a member of a club and working part time independent? Justify with probabilities.

8. A hospital runs a test to determine whether patients have certain disease. The test is not always accurate. The two-way table below summarizes the numbers of patients in the past year that received each result. Assume a patient is randomly selected.



a. What is the probability that the patient has the disease?
b. A patient from this group received a positive test result. What is the probability that he or she has the disease?
c. A patient from this group has the disease. What is the probability that he or she received a positive result on the test?
d. A "*false positive*" is when a patient receives a positive result on the test, but does not actually have the disease. What is the probability of a false positive?
e. What is the probability that a patient receives accurate test results?

Answers

You Try It $5.4.A$

a. $\frac{48}{81} \approx 59\%$
b. $\frac{5}{81} \approx 6\%$
c. $\frac{51}{81} \approx 63\%$
d. $\frac{20}{39} \approx 51\%$

Exercise $5.4.B$

Not independent because $P(\text{two or more credit cards}) \neq P(\text{two or more credit cards} \mid \text{male})$

$P(\text{two or more credit cards}) = \frac{39}{81} \approx 0.4815$

$P(\text{two or more credit cards} \mid \text{male}) = \frac{19}{33} \approx 0.5758$

This page titled 5.4: Contingency Tables, Conditional Probability, and Independence is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by Andy Jones.

- 12.7: Exercises by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

5.5: Odds

Learning Objectives

- Determine odds for and against an event.
- Convert between probability and odds.

Introduction

Up to this point, we have used probability to describe how likely an event E is by comparing the number of outcomes in the event to the number of outcomes in the entire sample space. For example, when flipping a fair coin, the probability of getting tails is

$$P(\text{tails}) = \tfrac{1}{2} = 50\%.$$

This tells us that half of all possible outcomes result in tails.

In everyday language, though, people often describe chance in a different way. You may hear someone say the chance of getting tails is "fifty-fifty" or "1 to 1." These expressions compare the likelihood of tails directly to the likelihood of heads. This comparison is not a probability -- it is an example of *odds*.

Probability compares an event to the entire sample space. Odds compare an event to its complement. This distinction is essential as we move into the next section, where we explore how odds are defined, how they differ from probability, and how to convert between the two.

Odds

Odds provide a second way to describe how likely an event is, and they differ fundamentally from probability. Probability compares an event to the entire sample space, while odds compare an event only to its complement -- the outcomes in which the event does *not* occur.

Odds measure how many outcomes favor an event versus how many do not. For an event E, **odds in favor** of E can be written as

$$\text{number of outcomes in } E : \text{number of outcomes in } E'$$

where E' is the complement of E.

In contrast, probability of E is written as

$$\frac{\text{number of outcomes in } E}{\text{number of outcomes in the sample space}}.$$

Because odds and probability use different comparisons, we will write them differently to avoid confusion. Odds are typically expressed as a **ratio**, such as $1:5$, read "one to five." Probability is usually expressed as a **fraction** such as $\frac{1}{5}$, read "one in five." Of course, probability can also be expressed as a decimal or percentage.

Understanding the distinction between odds and probability is essential, because each is used in different contexts -- probability in statistical reasoning and odds in settings such as gambling, risk assessment, and everyday descriptions of chance.

Example 5.5.1

You roll a fair six-sided die Find the odds in favor of getting a 3. Then, calculate the probability of getting a 3.

Solution

The key idea is that odds compare the number of outcomes in the event to the number of outcomes not in the event, while probability compares the event to the entire sample space.

The sample space has 6 outcomes:

- 1 outcome is a 3
- 5 outcomes are not a 3

The odds in favor of getting a 3 are $1:5$. This is read as "one to five."

The probability of getting a 3 is $\frac{1}{6}$. This is read as "one in six."

Odds and probability describe the same event but express its likelihood in different ways - one as a ratio comparing an event to its complement, the other as a fraction comparing the event to the entire sample space.

Example 5.5.2

Determine the probability and the odds in favor of rolling a sum of 5 with two fair six-sided dice.

Solution

To solve this, first identify all outcomes that produce the event and compare them either to the entire sample space (for probability) or to the complement (for odds).

There are four ways to roll a sum of 5: (1, 4), (2, 3), (3, 2), (4, 1).

Since each die has 6 faces, the sample space contains $6 \times 6 = 36$ possible ordered pairs.

The probability of rolling a sum of 5 is $\frac{4}{36} = \frac{1}{9}$. This is read as "one in nine."

If 4 outcomes produce a sum of 5, then $36 - 4 = 32$ outcomes do not produce a sum of 5.

Thus, the odds in favor of rolling a sum of 5 are $4:32$ which simplifies to $1:8$. This is read as "one to eight."

Example 5.5.3

A jewelry box contains 5 white pearl rings, 2 gold rings, and 6 silver rings. What are the odds in favor of drawing a silver ring?

Solution

To find odds, compare the number of outcomes in the event to the number of outcomes not in the event.

There are 13 rings in the jewelry box.

- Outcomes in the event "drawing a silver ring": 6
- Outcomes in the complement "not drawing a silver ring": $5 + 2 = 7$

Therefore, the odds in favor of drawing a silver ring are $6:7$. This ratio is read as "six to seven," meaning that for every 6 favorable outcomes, there are 7 unfavorable ones.

You Try It 5.5.A

A shelf contains 8 mystery novels, 3 science-fiction novels, and 4 historical novels. If one book is selected at random, what are the probability and odds in favor of choosing a mystery novel?

Answer available at the end of this section.

Odds Against

In many real-world settings -- especially sports, gaming, and gambling -- odds are often expressed as **odds against** rather than odds in favor. While odds in favor compare the number of favorable outcomes to the number of unfavorable outcomes, odds against reverse this comparison.

Odds against an event E are defined as

$$\text{number of outcomes in } E' : \text{number of outcomes in } E$$

where E' is the complement of E.

Because odds against simply reverse the order of the ratio, they are the *reciprocal* of the odds in favor. For example, if the odds in favor of an event are $3 : 7$, then the odds against the event are $7 : 3$.

It is essential to specify whether odds are being stated in favor or against, because the meaning changes depending on the direction of the comparison. In probability contexts, both forms appear, but in gambling and sports contexts, odds against are far more common.

Example 5.5.4

You draw a card from a standard deck of 52 cards. Find the odds against drawing a diamond.

Solution

One-fourth of the deck (or 13 cards) are diamonds and the remaining $52 - 13 = 39$ cards are not diamonds. The ratio of unfavorable outcomes (39) to favorable outcomes (13) is $39 : 13$, which simplifies to $3 : 1$.

Example 5.5.5

Roulette wheels contain 38 slots: 0, 00, and the numbers 1 through 36.

- The slots 0 and 00 are green.
- The even-numbered slots are red.
- The odd-numbered slots are black.

When the wheel is spun and the marble is released, it eventually falls into one slot. Find the odds against the marble landing in a red slot.

Solution

To compute odds against an event, compare the number of unfavorable outcomes to the number of favorable outcomes.

There are 18 even numbers from 1 to 36 so there are 18 red slots (favorable outcomes). There are $38 - 18 = 20$ remaining slots (the odd numbers, 0, and 00.) These are unfavorable outcomes.

The odds against landing in a red slot are $20 : 18$ which simplifies to $10 : 9$.

You Try It $5.5.B$

A bag contains 6 red marbles, 3 blue marbles, and 1 yellow marble.

a. What are the odds in favor of choosing a red marble?
b. What are the odds against choosing a blue marble?

Answer available at the end of this section.

Converting from Odds to Probability

Probability and odds both measure how likely an event is, but they do so in different ways. Probability represents the fraction of all possible outcomes in which the event occurs. Odds, on the other hand, compare the number of favorable outcomes directly to the number of unfavorable outcomes.

For example, imagine a clinical trial reporting that a new drug has odds of 5 to 1 in favor of curing a disease. This means there are 5 successes (cures) for every 1 failure (no cure).

To convert this to a probability, we compare the number of successes to the total number of outcomes:

$$P(\text{cure}) = \frac{5}{5+1} = \frac{5}{6}.$$

You can use this same process to convert chance expressed in odds to chance expressed as probability.

Converting from Odds to Probability

If the odds in favor of an event are $A : B$, then $P(A) = \dfrac{A}{A+B}$.

If the odds against an event are $A : B$, then $P(A) = \dfrac{B}{A+B}$.

Example $5.5.6$

The odds in favor of a shipment arriving on time are $9 : 4$. What is the probability that the shipment arrives on schedule?

Solution

In this scenario, the ratio $9 : 4$ means there are 9 outcomes where the shipment arrives on time and 4 outcomes where it doesn't arrive on time.

The total number of possible outcomes represented by the odds is $9 + 4 = 13$.

The probability that the shipment arrives on time is $\frac{9}{9+4} = \frac{9}{13}$.

Example 5.5.7

The odds against Kendrick winning the tennis match are $5:2$. What is the probability that Kendrick wins?

Solution

In this scenario, the ratio $5:2$ means there are 5 outcomes where Kendrick loses and 2 outcomes where Kendrick wins.

The total number of possible outcomes represented by the odds is $5+2=7$.

The probability that Kendrick wins is $\frac{2}{5+2}=\frac{2}{7}$.

You Try It 5.5.C

Jenny estimates her odds of earning an A on the next test as $7:18$. What is Jenny's probability of earning an A?

Answer available at the end of this section.

Converting from Probability to Odds

In many situations, chances are expressed as a probability, but you may want to describe them using odds instead.

The probability of an event is $\frac{3}{8}$. We want to convert from probability written as a fraction to odds using a ratio.

A probability of $\frac{3}{8}$ means 3 parts represent successful outcomes while the remaining $8-3=5$ parts represents unsuccessful outcomes.

Thus, the odds in favor of the event are $3:5$.

The same type of reasoning can be applied to covert any fractional probability to odds as seen in the following examples.

Example 5.5.8

A report states that a new air-filtering system has a probability $\frac{4}{5}$ of removing a harmful contaminant during testing.

What are the odds in favor of the system removing the contaminant?

Solution

A probability of $\frac{4}{5}$ means 4 parts represent successful outcomes while the remaining $5-4=1$ part represents unsuccessful outcomes.

Thus, the odds in favor of the system removing the contaminant are $4:1$.

Example 5.5.9

The weather forecast states that the probability of rain today is 40%. What are the odds in favor of rain?

Solution

The probability of 40% can be written as a fraction: $\frac{40}{100}=\frac{2}{5}$.

This means that out of 5 possible outcomes represented by the probability 2 outcomes are favorable (rain) and $5 - 2 = 3$ outcomes are unfavorable.

Therefore, the odds in favor of rain are $2 : 3$

There are other valid approaches to express the probability directly as a ratio of favorable to unfavorable outcomes.

For example, we could write $40\% : 60\%$ which simplifies to $2 : 3$.

We could also first form a ratio with decimals such as $0.40 : 0.60$. Scale each part by 100 to get $40 : 60$ and then simplify by a factor of 20 to get $2 : 3$.

However, no matter the methods, odds must be written as a ratio of two whole numbers. There should be no decimals or percentages left in odds.

You Try It $5.5.D$

The probability that someone who takes a flu shot stills gets the flu is 12%. What are the odds in favor of getting the flu if you get the shot?

Answer available at the end of this section.

While reasoning along these lines can be used to convert any probability to odds, it may be helpful to have a formula -- especially when probability is given as a decimal.

Converting from Probability to Odds

If the probability of event E is p, then the odds in favor of event E are $p : (1 - p)$.

If the probability of event E is p, then the odds against event E are $(1 - p) : p$.

Example $5.5.10$

The probability winning a prize in a game is 0.15. What are the odds against winning a prize?

Solution

Use the formula to convert probability to odds against, where $p = 0.15$:

$$(1 - p) : p$$

$$(1 - 0.15) : 0.15$$

$$0.85 : 0.15$$

Odds are written using only whole numbers, so we scale both parts by 100:

$$85 : 15.$$

Finally, simplify the ratio by dividing both numbers by 5:

$$17 : 3$$

You Try It $5.5.E$

The probability that it snows tomorrow is 70%. What are the odds against snow tomorrow?

Answer available at the end of this section.

Section Summary

Odds provide another way to describe the likelihood of an event. Instead of focusing on the proportion of successful outcomes as probability does, odds compare the number of outcomes in an event directly to the number of outcomes in its complement.

The direction of the comparison determines whether the odds are in favor of the event or against it:

- Odds in favor of event A are the ratio of the number of outcomes in A to the number of outcomes in A'.
- Odds against event A are the ratio of the number of outcomes in A' to the number of outcomes in A.

Because the order of the comparison changes the meaning, it is essential to specify whether odds are in favor of the event or against it.

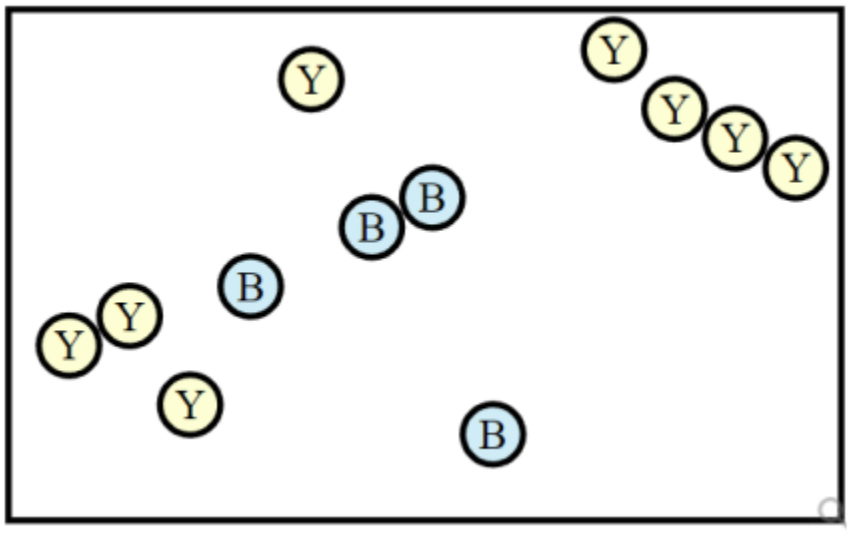
Odds and probability are closely related, and it is often useful to convert between them.

- If the odds in favor of an event are $A : B$, then the probability the event occurs is $\frac{A}{A+B}$.
- If the odds against the event are $A : B$, them the probability that the event occurs is $\frac{B}{A+B}$.
- If the probability of an event is p, then the odds in favor of the event are $p : (1 - p)$ and the odds against the event are $(1 - p)$: p.

These relationships allow you to move easily between the two ways of expressing chance, depending on which form is more convenient for interpretation or calculation.

Section 5.5 Review Questions

1. Peter has 12 marbles as shown in the diagram below. Each marble is either blue (labeled with a B) or yellow (labeled with a Y). Peter chooses one of the marbles at random. Determine the following probabilities as fractions and odds as ratios in simplest form.



   a. What is the probability that Peter chooses a yellow marble?
   b. What are the odds in favor of Peter choosing a yellow marble?

c. What are the odds against Peter choosing a yellow marble?

2. Eleven poker chips are placed in a jar. Ten of them are numbered consecutively from 1 through 10, and the eleventh chip is numbered a 6. A chip is drawn at random. Find the odds in favor of drawing a 6 from the jar.

3. The game board on *Jeopardy* is divided into 30 squares (there are six categories with five answers in each category). In the Double Jeopardy round, 2 Daily Double squares are hidden among the 30 squares. What are the odds in favor of randomly choosing a Daily Double square on the first selection?

4. Dave's wardrobe contains 29 T-shirts, of which 15 are gray. Dave selects a T-shirt at random.
   a. What is the probability that Dave selects a gray T-shirt?
   b. What are the odds in favor of Dave selecting a gray T-shirt?

5. A door prize at a party with 25 people is given to a guest choosing the lucky paper plate from plates numbered 1 to 25. What are the odds against winning the door prize?

6. A bag contains 6 red marbles, 3 blue marbles, and 1 yellow marble.
   a. What is the probability of choosing a red marble?
   b. What are the odds of choosing a red marble?
   c. What are the odds against choosing a blue marble?

7. Kendra has various cereals on a shelf in the cabinet. Five of the cereals contain corn, two contain rice, and four contain oats. Without looking, she selects a box of cereal for breakfast.
   a. What are the odds in favor of selecting a cereal that contains oats?
   b. What are the odds against selecting a cereal that contains rice?
   c. What is the probability of selecting a cereal that contains corn?

8. Answer these questions relating odds and probability:
   a. If the odds of winning a game are $9:7$, what is the probability of losing the game?
   b. If the probability of event E is $\frac{2}{13}$, what are the odds in favor of event E?
   c. A meteorologist estimates the chance of snow tomorrow as 75%. What are the odds in favor of snow?

9. Experts calculate the probability of a particular driver winning a NASCAR race to be 0.3. What are the odds against the driver winning the race?

10. Answer the questions relating probability to odds.

   a. The probability that a person will get the flu in a given year is 11%. What are the odds in favor of a person getting the flu?
   b. The probability winning a prize in a game is 0.09. What are the odds against winning a prize?
   c. The odds in favor of your favorite sports team winning a game are $24:7$. What is the probability they won't win?
   d. The odds in favor of a tornado watch being issued are $37:86$. What is the probability of a tornado watch being issued?

Answers

You Try It $5.5.A$

The probability of choosing a mystery novel is $\frac{8}{15}$.

The odds in favor of choosing a mystery novel are $8:7$.

You Try It $5.5.B$

a. $3:2$
b. $7:3$

You Try It $5.5.C$

$\frac{7}{25} = 0.28$, or 28%.

You Try It $5.5.D$

$3:22$

You Try It $5.5.E$

$3:7$

This page titled 5.5: Odds is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

6: PROBABILITY RULES AND EXPECTATION

6.1: Fundamental Properties of Probability
6.2: The Additive Rules of Probability
6.3: The Multiplicative Rules of Probability
6.4: Tree Diagrams and Multistage Probability
6.5: Expectation Involving Chance

This page titled 6: Probability Rules and Expectation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

6.1: Fundamental Properties of Probability

Learning Objectives

- Recognize legitimate probability values.
- Define the complement of an event.
- Calculate probability using a complement.

Introduction

As we saw in Chapter 5, probability can be computed by listing all possible outcomes in a sample space and identifying which of those outcomes belong to the event of interest. The sample space was sometimes given as a listed set of outcomes, in a Venn diagram, or in a contingency table. While this approach is useful for simple situations, it quickly becomes impractical when the number of outcomes is large or when events are defined in more abstract ways. For this reason, it is essential to learn the formal **rules of probability**, which allow you to compute probabilities efficiently and accurately without relying on a full sample space every time.

In this section, you will be introduced to some foundational properties of probability. You will also see how the complement of an event plays a central role in these rules, and you will learn to apply the *Complement Rule* to find the probability of an event by working with its complement instead. This approach is especially helpful when the complement is easier to describe or count than the event itself.

Two Fundamental Properties of Probability

When using the theoretical probability formula

$$P(A) = \frac{\text{number of outcomes in event } A}{\text{number of outcomes in the sample space } S},$$

two fundamental restrictions apply:

- The number of outcomes in event A must be greater than or equal to 0.
- The number of outcomes in the sample space S must be greater than 0.

These restrictions guarantee that any probability value is never negative. Furthermore, the number of outcomes in event A can never exceed the total number of outcomes in the sample space S. Therefore, the probability ratio can never be greater than 1.

Together, these observations give the most basic property of probability: every probability lies between 0 and 1, inclusive.

> **Property 1: Bounds on Probability**
>
> For any event A, $0 \le P(A) \le 1$.

If $P(A) = 0$, the event is **impossible**. If $P(A) = 1$, the event is **certain**.

This property reflects the fact that probability measures how likely an event is, with 0 and 1 representing the two extreme cases.

Example 6.1.1

Which statements use valid values for probability?

a. A researcher claims that the probability a new medication is effective is 1.08.
b. A survey reports that the probability a randomly selected person prefers tea over coffee is 0.65.
c. A report states that the probability a randomly selected plant survives its first year is $\frac{9}{5}$.
d. A weather model predicts that the probability of snowfall tomorrow is 0.

Solution

Only statements b and d are valid because they both use probability values that lie between 0 and 1 inclusive. In both statements a and c, the stated probabilities are more than 1.

You Try It $6.1.A$

Which statements use valid values for probability?

a. A nutrition study claims that the probability a child chooses a fruit snack over a cookie is -0.12.
b. A transportation report states that the probability a bus arrives on time is 0.81.
c. A wildlife survey reports that the probability a tagged bird returns to its nesting site is $\frac{11}{10}$.
d. A classroom poll finds that the probability a student completes the homework is 112%.

Answer available at the end of this section.

Property 2: Total Probability of the Sample Space

The probabilities of all simple outcomes in a sample space add to 1. Formally,

$$\sum_{i=1}^{n} P(A) = 1$$

when the sum is taken over all n simple outcomes of the sample space.

This property expresses the certainty that *some* outcome in the sample space must occur. For example, when tossing a coin, the sample space is $\{H, T\}$. Each simple outcome has probability 0.5. Summing these probabilities gives $0.5 + 0.5 = 1$, which reflects the fact that every coin toss is certain to result in either heads or tails.

Example $6.1.2$

A spinner has three equal sections labeled A, B, and C. The probabilities of landing on A and B are $P(A) = 0.3$ and $P(B) = 0.45$.

Find $P(C)$.

Solution

Using Property 2, we know the sum of the probabilities must be 1. Substitute known values and solve for $P(C)$.

$$\begin{aligned} P(A) + P(B) + P(C) &= 1 \\ 0.3 + 0.45 + P(C) &= 1 \\ 0.75 + P(C) &= 1 \\ P(C) &= 1 - 0.75 = 0.25 \end{aligned}$$

You Try It 6.1.B

A probability model for a four-outcome experiment lists the following values:

$$P(A) = 0.15 \; P(B) = 0.4 \; P(C) = 0.32 \; P(D) = 0.18$$

Is this a valid probability model? Explain why or why not.

Answer available at the end of this section.

Complement Rule of Probability

In probability, it is often useful to consider not only an event itself but also everything that does not belong to that event. This idea was captured by the *complement* of an event as we saw in Section 5.3. Understanding complements allows us to compute probabilities more efficiently, especially when the event of interest is complicated but its complement is simple to describe.

Because event A and its complement A' together form the entire sample space, the probabilities of these two events must sum to 1. This leads to the **Complement Rule**:

Complement Rule of Probability

For any event A,

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

This rule can be rearranged to solve for either probability:

$$P(A) = 1 - P(A') \; \mathrm{P}(A') = 1 - P(A)$$

These relationships are extremely useful. In many situations, it is much easier to compute the probability of the complement A' than the probability of the event A itself. Once $P(A'$ is known, the probability of the event follows immediately from $P(A) = 1 - P(A')$.

Example 6.1.3

Suppose the probability of getting the flu this winter is 0.43. What is the probability that you will not get the flu?

Solution

Let event A be "get the flu."

We are given $P(A) = 0.43$. The event "do not get the flu" is A'.

$$\begin{aligned} P(A') &= 1 - P(A) \\ &= 1 - 0.43 \\ &= 0.57 \end{aligned}$$

The probability that you will not get the flu is 0.57.

Example 6.1.4

The probability that a randomly selected email is spam is $\frac{1}{6}$.

Find the probability that an email is not spam.

Solution.

Let event A be "email is spam."

We are given $P(A) = \frac{1}{6}$. The event "email is not spam" is A'.

$$\begin{aligned} P(A') &= 1 - P(A) \\ &= 1 - \frac{1}{6} \\ &= \frac{6}{6} - \frac{1}{6} \\ &= \frac{5}{6} \end{aligned} .$$

The probability the e-mail is not spam is $\frac{5}{6}$.

Example 6.1.5

A medical study reports that 18% of patients don't respond to a particular treatment.

Find the probability that a patient does respond to the treatment.

Let A be the event “a patient responds to the treatment.”

We are given that $P(A') = 18\% = 0.18$. Apply the Complement Rule:

$$\begin{aligned} P(A) &= 1 - P(A') \\ &= 1 - 0.18 \\ &= 0.82 \end{aligned} .$$

The probability that a patient responds to the treatment is 0.82 (or 82%).

You Try It 6.1.C

A manufacturing report shows that the probability a randomly selected part fails inspection is $\frac{3}{20}$.

Find the probability that the part passes inspection.

Answer available at the end of this section.

Example 6.1.6

A streaming service tracks whether users watch a comedy or drama during a given week. The service reports that the probability a randomly selected user watches a comedy or drama is 0.78.

Find the probability that a user watches neither a comedy nor a drama during the week.

Solution

The event "watches neither a comedy nor a drama" is the complement of "watches a comedy or a drama."

$$\begin{aligned} P(A \text{ or } B)' &= 1 - P(A \text{ or } B) \\ &= 1 - 0.78 \\ &= 0.22 \end{aligned}$$

Thus, the probability that a user watches neither a comedy nor a drama is 0.22.

You Try It $6.1.D$

A community center offers two weekend classes: Art and Music. A survey shows that the probability a randomly selected resident signs up for neither Art nor Music is $\frac{3}{16}$.

Find the probability that a resident signs up for at least one of the classes.

Answer available at the end of this section.

Section Summary

Two fundamental probability properties describe how probabilities behave within a sample space:

1. For any event A, the probability must lie between 0 and 1: $0 \leq P(A) \leq 1$.

   A probability of 0 represents an impossible event, while a probability of 1 represents a certain event.

2. The probabilities of all simple outcomes in a sample space must sum to 1: $\sum_{i=1}^{n} P(A) = 1$

   This reflects the fact that one of the outcomes in the sample space must occur.

The complement of an event A, written A', consists of all outcomes in the sample space that are not in A. Because an event and its complement together make up the entire sample space, their probabilities must add to 1. This relationship is known as the *Complement Rule*:

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

The rule can be rearranged to solve for either probability:

$$P(A) = 1 - P(A') \text{ and } P(A') = 1 - P(A).$$

Section 6.1 Review Questions

1. A probability experiment is conducted. Which of these *cannot* be considered the probability of an event?
   a. 0
   b. -0.3
   c. 34
   d. 125%
   e. $\frac{4}{3}$
   f. 0.56
   g. 1
   h. 1.06
   i. 12.5%

2. If the probability that a person lives in an industrialized country of the world is $\frac{1}{5}$, find the probability that a person does not

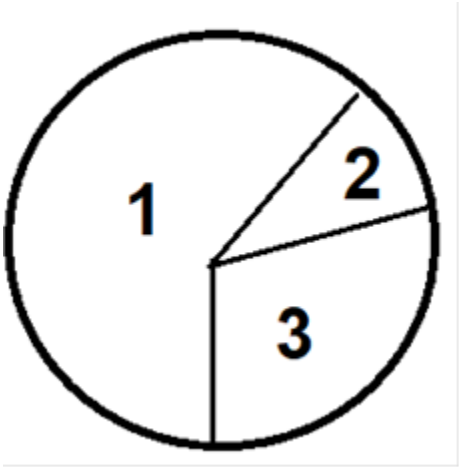
live in an industrialized country.

3. You pull a random card from a deck of playing cards. Use the Complement Rule of Probability to find the probability it is not a heart.

4. According to *Time* magazine, 69% of adults favor gun licensing. If you choose one adult at random, what is the probability that the selected adult doesn't favor gun licensing?

5. A large bucket contains only red, white, and blue blocks. One block will be randomly selected. Which of the these *cannot* describe the probabilities of getting each color of block?
   a. P(red) = 0.3, P(white) = 0.3, P(blue) = 0.3
   b. P(red) = $\frac{1}{2}$, P(white) = $\frac{1}{3}$, P(blue) = $\frac{1}{6}$
   c. P(red) = $\frac{1}{20}$, P(white) = $\frac{3}{5}$, P(blue) = $\frac{9}{20}$
   d. P(red) = 0.75, P(white) = 0.4, P(blue) = -0.15
   e. P(red) = 0.01, P(white) = 0.70, P(blue) = 0.29

6. On the spinner shown below, the probability of spinning a "1" is 58% and the probability of spinning a "2" is 9%. What is the probability of spinning a 3?



7. A school club offers two activities during lunch: Chess and Robotics.. A student survey shows that there is a 72% chance that a student participates in either Chess or Robotics. Find the probability that a student will participate in neither activity.

8. Sam knows the chance he can make it home without hitting any red lights is 0.15. What is the chance that Sam will hit at at least one red light on the way home?

Answers

You Try It $6.1.A$

b only

You Try It $6.1.B$

No. The sum of the probabilities is not 1. It is 1.05.

You Try It $6.1.C$

$\frac{17}{20}$

You Try It $6.1.D$

$\frac{13}{16}$

This page titled 6.1: Fundamental Properties of Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

6.2: The Additive Rules of Probability

Learning Objectives

- Calculate probabilities using the *Additive Rule of Probability* for events that are mutually exclusive.
- Calculate probabilities using the *Additive Rule of Probability* for events that are not mutually exclusive.
- Identify whether two events are mutually exclusive.

Introduction

In this section we calculate the probability that event A occurs or event B occurs. Many real situations involve overlapping events, and rather than listing or analyzing every outcome in the sample space, we use additive probability rules to combine the probabilities of two events efficiently. These rules allow us to determine $P(A \text{ or } B)$ whether the events overlap or do not overlap. As you move through this section, you will see how these formulas simplify problems and help you handle "or" statements with confidence.

Meaning of "*Or*"

In probability, the word **"or"** has a precise mathematical meaning that is broader than how we often use it in everyday language. In everyday conversation, "or" can sound exclusive - as if only one choice can happen. In probability, however, "or" is inclusive. This means that A or B refers to every outcome in which event A occurs, event B occurs, or both events occur together. In set notation, this combined event $(A \text{ or } B)$ is written as the union $A \cup B$.

Thinking of "or" as inclusive ensures that we capture all outcomes where *at least one* of the two events happens. If a student is selected who plays soccer or basketball, this includes students who play soccer only, basketball only, and students who play both. If a shopper buys fruit or vegetables, the shopper who buys both is still part of the event. This inclusive meaning is essential because it aligns with how sample spaces are structured: any outcome belonging to either event must be counted when determining the probability of their union.

This interpretation of "or" is the foundation for the *Additive Rules of Probability*. In every case, the inclusive meaning of "or" ensures that $P(A \text{ or } B)$ accurately represents the probability that at least one of the two events occurs.

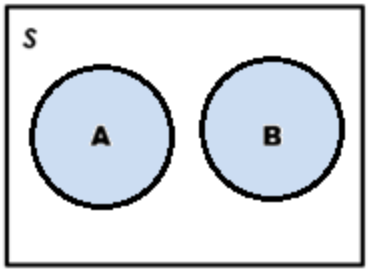
The Additive Rules for 'Or' Events

When the probabilities of two events are known, we can use these probabilities to calculate the probability of their union -- the event that at least one of them occurs. And so $P(A \text{ or } B)$ is equivalent to $P(A \cup B)$.

We use the **Additive Rules of Probability** to find it. The development of these rules depend on how the two events relate to each other within the sample space.

Mutually Exclusive Events

Two events are **mutually exclusive** (also called **disjoint**) when they cannot occur at the same time. Their outcome sets do not overlap as shown in the Venn diagram below:



A or B

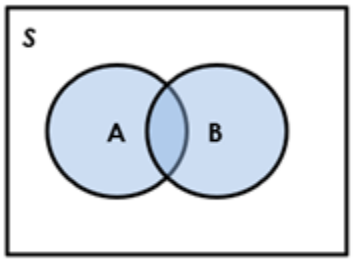
Because the events are completely separate, the probability that either event occurs is simply the sum of their individual probabilities:

$$P(A \cup B) = P(A) + P(B).$$

This works because each probability represents a distinct portion of the sample space. No outcome belongs to both events, so nothing is counted twice. For example, a single roll of a die cannot be both a 2 and a 5, so the probability of rolling a 2 or a 5 is just the sum of the two individual probabilities.

Non-Mutually Exclusive Events

However, many real-world events are **not mutually exclusive** (or **not disjoint**) and can occur together.



A or B

When events overlap, simply adding their probabilities double-counts the outcomes that belong to both events. To correct this, we subtract the probability of the intersection:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

This formula ensures that every outcome in the union is counted exactly once. The overlap $A \cap B$ is included in both $P(A)$ and $P(B)$, so subtracting it removes the extra count. For example, if some students play both soccer and basketball, those students should be counted only once when determining the probability that a student plays soccer or basketball.

Together, the rules for mutually exclusive and those that are not mutually exclusive events are known as the *Additive Rules of Probability*.

> **Additive Rules of Probability**
>
> For events A and B,
>
> $P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive.

$P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$ if A and B are not mutually exclusive.

Example 6.2.1

A raffle drum contains 50 tickets. Of these, 14 are labeled Prize A, 9 are labeled Prize B, and the remaining tickets are labeled No Prize. One ticket is drawn at random.

What is the probability that the ticket drawn is Prize A or Prize B?

Solution

Define the events:

A is Prize A and B is Prize B.

We want to find $P(A \cup B)$ and these events are mutually exclusive because a single ticket cannot be labeled both Prize A and Prize B.

- There are 14 tickets labeled Prize A, so $P(A) = \frac{14}{50}$.
- There are 9 tickets labeled Prize B, $P(B) = \frac{9}{50}$.

To find $P(A \cup B)$, apply the Additive Rule of Probability for events that are mutually exclusive:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ &= \frac{14}{50} + \frac{9}{50} \\ &= \frac{23}{50} \end{aligned}$$

The probability of drawing Prize A or Prize B is $\frac{23}{50}$ or 46%.

Example 6.2.2

In a certain voting precinct, 39% of voters are Republicans, 35% are Democrats, and 26% are Independents. If a voter is selected at random, find the probability that the voter is a Democrat or an Independent.

Solution

Define the events:

A is Democrat and B is Independent.

We want to find $P(A \cup B)$ and these events are mutually exclusive because a voter can belong to only one political party.

As stated in the problem, $P(A) = 0.35$ and $P(B) = 0.26$.

To find $P(A \text{ or } B)$, apply the Additive Rule of Probability for events that are mutually exclusive:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ &= 0.35 + 0.26 \\ &= 0.61 \end{aligned}$$

Thus, there is a 61% chance of selecting a voter who is a Democrat or an Independent.

You Try It 6.2.A

A library has 120 books on a display table. Of these, 45 are fiction, 30 are biographies, and the remaining books belong to other categories. One book is selected at random.

What is the probability that the selected book is fiction or a biography?

Answer available at the end of this section.

Example 6.2.3

Consider the experiment of randomly selecting a card from a standard deck of 52 playing cards.

What is the probability that the card selected is a spade or a face card?

Solution

Define the events:

A is the card is a spade and B is the card is a face card.

We want to find $P(A \cup B)$ and these events are not mutually exclusive. There are some cards that are both a spade and a face card. Refer to Section 5.1 for an illustration of a deck of cards if needed.

In a standard deck,

- There are 13 spades and $P(A) = \frac{13}{52}$.
- There are 12 face cards (Jacks, Queens, Kings) and $P(B) = \frac{12}{52}$.
- There are 3 cards that are both spades and face cards (Jack, Queen, King of spades) and $P(A \cap B) = \frac{3}{52}$.

To find $P(A \cup B)$, apply the Additive Rule of Probability for events that are not mutually exclusive:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= \frac{22}{52} \\ &= \frac{11}{26} \approx 0.4231 \ (42.31\%) \end{aligned}$$

We subtract $\frac{3}{52}$ because the three cards that are both spades and face cards would otherwise be counted twice -- once in $P(A)$ and once in $P(B)$. Subtracting the intersection ensures each card is counted exactly once in the union.

You Try It 6.2.B

A bag contains 40 game tokens. Of these, 12 are red, 15 are numbered, and 5 are both red and numbered. One token is selected at random. What is the probability that the token is red or numbered?

Answer available at the end of this section.

Example 6.2.4

Crime statistics show 84% of the people arrested last year were males, 18% were under the age of 18, and 14% were both male and under 18. What is the probability that a randomly selected person from those arrested was male or under the age of 18?

Solution

Define the events:

A is male and B is under the age of 18.

We want to find $P(A \cup B)$ and these events are not mutually exclusive because there is a chance that a person arrested is both male and under the age of 18.

As stated in the problem, $P(A) = 0.84$, $P(B) = 0.18$, and $P(A \cap B) = 0.14$.

To find $P(A \cup B)$, apply the Additive Rule of Probability for events that are not mutually exclusive:

$$\begin{aligned} P(A\cup &= P(A) + P(B) - P(A \cap B) \\ &= 0.84 + 0.18 - 0.14 \\ &= 0.88 \end{aligned}$$

This means that 88% of the people arrested last year were male, under 18, or both. The subtraction of the overlap ensures that individuals who fall into both categories are counted only once.

Example 6.2.5

There is a 90% chance of snow, a 20% chance of strong winds, and a 15% chance of experiencing both snow and strong winds. What is the probability of snow or strong winds?

Solution

Define the events:

A is snow and B is strong winds.

We want to find $P(A \cup B)$ and these events are not mutually exclusive because there is a chance there will be both.

As stated in the problem, $P(A) = 0.90$, $P(B) = 0.20$, and $P(A \cap B) = 0.15$.

To find $P(A \cup B)$, apply the Additive Rule of Probability for events that are not mutually exclusive:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.90 + 0.20 - 0.15 \\ &= 0.95 \end{aligned}$$

There is a 95% chance of experiencing snow, strong winds, or both.

This subtraction step ensures that the 15% of days with both snow and strong winds are not counted twice -- once in $P(A)$ and once in $P(B)$.

You Try It 6.2.C

Suppose that 48% of employees in a company participate in the wellness program, 35% participate in the nutrition program, and 22% participate in both programs. If an employee is selected at random, what is the probability that the employee participates in the wellness program or the nutrition program?

Answer available at the end of this section.

Using an Additive Rule to Solve for Another Probability

Up to this point, we have used the *Additive Rules of Probability* to find the probability of the union of two events that are not mutually exclusive, $P(A \cup B)$. Once we know the probabilities of the individual events and their intersection, the additive rule allows us to combine them:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

However, this relationship works both ways. If we are given the probability of the union along with other probabilities, we can rearrange the additive rule to solve for any missing quantity. For example, if $P(A \cup B)$, $P(A)$, and $P(B)$ are known, we can solve for the intersection:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

Similarly, if $P(A \cup B)$, $P(B)$, and $P(A \cap B)$ are known, we can solve for $P(A)$:

$$P(A) = P(A \cup B) - P(B) + P(A \cap B).$$

And if $P(A \cup B)$, $P(A)$, and $P(A \cap B)$ are known, we can solve for $P(B)$:

$$P(B) = P(A \cup B) - P(A) + P(A \cap B).$$

These rearrangements show that the additive rule is not only useful for finding the probability of “A or B” -- it also provides a way to determine any of the related probabilities when the others are known.

There is no need to memorize these rearrangements as distinct formulas -- just use your knowledge of algebra to solve the *Additive Rule of Probability* for the unknown probability.

Example 6.2.6

In a certain city, 60% of households have high-speed internet, 85% have high-speed internet or cable TV, and 25% have both high-speed internet and cable TV. What is the probability that a randomly selected household has cable TV?

Solution

Define the events:

A is high-speed internet and B is cable TV.

These events are not mutually exclusive because there is a chance a household has both services. As stated in the problem,

$$P(A) = 0.60,\ P(A \cup B) = 0.85, \text{ and } P(A \cap B) = 0.25.$$

To find $P(B)$, apply the Additive Rule of Probability for events that are not mutually exclusive:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$0.85 = 0.60 + P(B) - 0.25$$
$$0.85 = P(B) + 0.35$$
$$P(B) = 0.85 - 0.35$$
$$P(B) = 0.50$$

There is a 50% chance of a household in this city having cable TV.

Example 6.2.7

On any given night, the probability that Nick has a cookie for dessert is 10%, the probability that he has ice cream is 50%, and the probability that he has a cookie or ice cream is 55%. What is the probability that Nick has both a cookie and ice cream?

Solution

Define the events:

A is Nick has cookie and B is Nick has ice cream

These events are not mutually exclusive because there is a chance that Nick has both desserts. As stated in the problem,

$$P(A) = 0.10,\ P(B) = 0.50, \text{ and } P(A \cup = 0.55.$$

To find $P(A \cap B)$, apply the Additive Rule of Probability for events that are not mutually exclusive:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$0.55 = 0.10 + 0.50 - P(A \cap B)$$
$$0.55 = 0.60 - P(A \cap B)$$
$$P(A \cap B) = 0.60 - 0.55$$
$$P(A \cap B) = 0.05$$

There is a 5% chance that Nicks has cookie and ice cream.

You Try It 6.2.D

At a local gym, there is a 60% chance that a randomly selected member attends yoga, an 85% chance that the member attends yoga or pilates, and a 25% chance that the member attends both yoga and pilates.

What is the probability that the member attends pilates?

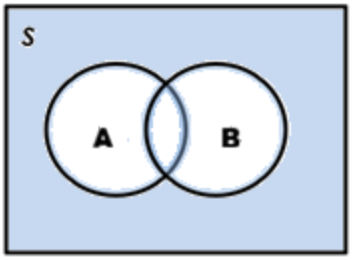
Answer available at the end of this section.

Using an Additive Rule to find Probability of Neither A nor B

In the previous examples, we have used one of the additive rules to find the probability of $P(A \cup B$ -- the probability that at least one of the events occurs. But sometimes the question asks for the probability that *neither* event occurs. Since the additive rules do not directly give $P(A \cup B)'$, we combine it with the *Complement Rule*, which states:

$$P(A \cup B)' = 1 - P(A \cup B).$$

As the Venn diagram illustrates, we can first use the additive rule to find $P(A \cup B)$, then subtract that value from 1 to obtain the probability of “neither.”



neither A nor B

Example 6.2.8

At a certain school, the probability that a male student plays football is 0.29, the probability that he plays basketball is 0.20, and the probability that he plays both sports is 0.04. What is the probability that a randomly selected male student plays neither sport?

Solution

Define the events:

A plays football and B plays basketball.

As stated in the problem,

$$P(A) = 0.29,\ P(B) = 0.20, \text{ and } P(A \cap B) = 0.04.$$

We want to find $P(A \cup B)'$, the probability a male student plays neither sport. Because the additive rule gives $P(A \cup B)$, we find that first:

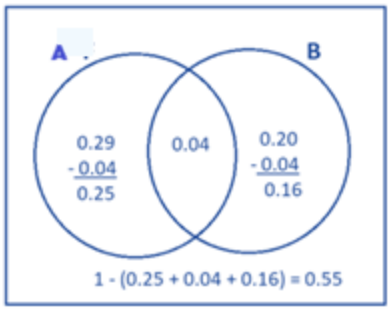
$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.29 + 0.20 - 0.04 \\ &= 0.45 \end{aligned}$$

Now, apply the *Complement Rule*:

$$\begin{aligned} P(A \cup B)' &= 1 - P(A \cup B) \\ &= 1 - 0.45 \\ &= 0.55 \end{aligned}$$

There is a 55% chance a male student plays neither football nor basketball.

You could also solve this problem by completing a Venn diagram, but the complement-plus-additive approach is typically faster and more general.



You Try It $6.2.E$

At a neighborhood pet adoption event, the probability that a visitor interacts with dogs is 38%, the probability that a visitor interacts with cats is 33%, and the probability that a visitor interacts with both dogs and cats is 12%.

What is the probability that a randomly selected visitor interacts with neither dogs nor cats?

Answer available at the end of this section.

Section Summary

The *Additive Rules of Probability* describe how to find the probability that **at least one** of two events occurs. For any events A and B, the probability of their union is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

This formula accounts for the fact that if the events overlap, the intersection $P(A \cap B)$ must be subtracted to avoid counting shared outcomes twice.

When the intersection contains no outcomes—meaning the events cannot occur together—the events are mutually exclusive (or disjoint). In that case, the intersection is zero, and a simpler additive rule can be used:

$$P(A \cup B) = P(A) + P(B).$$

In many applications, we are asked for the probability that neither event occurs. Because the additive rules only give the probability of at least one event occurring, we combine it with the *Complement Rule*. First compute the union using the additive rule, then subtract that value from 1 to obtain the probability of neither event.

$$P(A \cup B)' = 1 - P(A \cup B)$$

Section 6.2 Review Questions

1. At a community art studio, there is a 40% chance that a student will work on painting, a 70% chance that the student will work on sculpture, and a 24% chance that the student will work on both painting and sculpture. What is the probability that the student works on painting or sculpture?

2. At a weekend farmers market, a vendor offers fresh berries and homemade jam as add-ons to her fruit baskets. There is a 50% chance that a customer chooses both berries and jam, a 95% chance that the customer chooses berries or jam, and an 80% chance that the customer chooses berries. What is the probability that the customer chooses jam?

3. At a large music festival, the probability that an attendee buys a t-shirt is 0.65, and the probability that an attendee buys

a poster is 0.82. If the probability that an attendee buys both a t-shirt and a poster is 0.55, what is the probability that an attendee buys neither item?

4. At a weekend movie night, a group of friends votes on what to watch. The probability that they choose a comedy is 30%, and the probability that they choose an action movie is 60%. What is the probability that they choose a comedy or an action movie?

5. Among a group of boys, 70% like chocolate ice cream, 40% like strawberry ice cream, and 30% like both. If a boy is randomly selected from the group, what is the probability that he likes chocolate or strawberry ice cream?

6. The probability that a customer orders a burger or fries at a fast food restaurant is 80%. The probability that a customer orders a burger and fries is 50%. The probability that a customer orders fries is 60%. What is the probability that a customer orders a burger?

7. At Walton High School, 40% of athletes play football, 30% of athletes play basketball and 60% play football or basketball. What percent play football and basketball?

8. The table shows probabilities of the blood type of a randomly chosen person in the US.



Maria has Type B blood so she can receive transfusions from anyone with blood types O and B. What is the probability that a randomly chosen person can donate blood to Maria?

9. A jar contains 12 red marbles numbered 1 to 12, and 8 blue marbles numbered 1 to 8. A marble is drawn at random from the jar. Find the probability of each event.
   a. the marble is red
   b. the marble is odd-numbered
   c. the marble is red or odd-numbered
   d. the marble is blue or even-numbered

10. Determine whether each pair of events are mutually exclusive:
    a. You roll a single die.
       i. Event *A*: The result is an odd number.
       ii. Event *B*: The result is an even number.
    b. You roll a single die.
       i. Event *A*: The result is a number less than 4.
       ii. Event *B*: The result is an odd number.
    c. You select a single card from a deck of cards.
       i. Event *A*: The card is a club.
       ii. Event *B*: The card is a King.
    d. You select a student in your college:
       i. Event *A*: The student is a sophomore.
       ii. Event *B*: The student is a business major.

11. A single card is drawn from a deck. Find the probability of selecting the following:
    a. A club or a diamond
    b. A 4 or a diamond
    c. A Jack or a black card

12. In a subdivision, 67% of homes have garages, 32% have a patio, and 13% have both.
    a. Find P(a house has a garage or a patio).
    b. Find P(a house has neither a garage or a patio).
    c. Find P(a house has a garage but not a patio).
    d. Find P(a house has a patio given that it has a garage) .
    e. Find P(a house has a garage given that it has a patio).
    f. Are the events "*having a garage*" and "*having a patio*" mutually exclusive?

g. Are the events "*having a garage*" and "*having a patio*" independent?

13. Suppose that $P(A) = 0.32$, $P(B) = 0.48$, and $P(A \cup B) = 0.70$. Are events A and B mutually exclusive? Explain.

Answers

You Try It $6.2.A$

$\frac{75}{120} = \frac{5}{8} = 0.625 = 62.5\%$

You Try It $6.2.B$

$\frac{22}{40} = \frac{11}{20} = 0.55 = 55\%$

You Try It $6.2.C$

61%

You Try It $6.2.D$

50%

You Try It $6.2.E$

41%

This page titled 6.2: The Additive Rules of Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

6.3: The Multiplicative Rules of Probability

Learning Objectives

- Calculate probabilities using the *Multiplicative Rule of Probability* for two or more events that are not independent.
- Calculate probabilities using the *Multiplicative Rule of Probability* for two or more events that are independent.
- Use conditional probability to determine whether two events are independent.

Introduction

Section 6.2, we focused on the *Additive Rules of Probability,* which allow us to find the probability of $A \cup B$, or "A or B." That rule helped us determine how likely it is that at least one of two events occurs.

We now shift to the ***intersection*** of two events, written as $A \cap B$, or "A and B." Instead of asking how likely it is for one event or the other to occur, we ask how likely it is for both events to occur together. This leads us to the *Multiplicative Rules of Probability*, which describe how to compute the probability of an intersection in situations involving independent events and dependent events.

Just as the additive rules required us to understand unions and overlaps, the multiplicative rules require us to understand how events relate to one another -- whether one event affects the likelihood of the other.

The Multiplicative Rules for 'And' Events

For two events A and B, we may want to calculate calculate the probability of their intersection -- the event that both of them occurs. And so $P(A \text{ and } B)$ is equivalent to$P(A \cap B)$.

We use the **Multiplicative Rules of Probability** to find it. The key idea behind the development of these rules is that the probability of two events occurring together depends on how the second event behaves once the first event has already happened. This leads naturally to the concept of conditional probability as introduced in Section 5.4.

Dependent Events

We begin with the more general situation of **dependent events**. Two events are dependent when the occurrence of one event changes the probability of the other. This relationship is captured by **conditional probability**, written as $P(B|A)$, which means "the probability that B occurs given that A has already occurred."

Recall that conditional probability can be expressed using the formula

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

We can rearrange the formula for conditional probability to solve for $P(A \cap B)$ as follows:

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

Multiplying both sides of the equation by $P(A)$ gives

$$P(A) \cdot P(B|A) = P(A \cap B)$$

or

$$P(A \cap B) = P(A) \cdot P(B|A)$$

Multiplicative Rule for Dependent Events

For events A and B, $P(A \cap B) = P(A) \cdot P(B|A)$.

This formula works because first, we enter event A with probability $P(A)$. Then, *within that restricted situation*, we multiply by the probability that B occurs *given* that we are already inside A. This two-step structure -- *first A, then B within A* -- is what makes the rule multiplicative.

The next few examples illustrate finding $P(A \cap B)$ when event B depends on event A. We will discuss the case where the events are independent later in this section.

Example 6.3.1

In your drawer you have 10 loose socks, 6 of which are white. You reach in and randomly grab two socks.

What is the probability that both socks are white?

Solution

To find the probability that both socks selected are white, we treat the selection as a two-stage process and use the *Multiplicative Rule for Dependent Events*. Because the socks are drawn without replacement, the outcome of the second draw depends on what happened in the first draw.

Define the events:

W_1 is the first sock is white and W_2 is the second sock is white.

We want to find $P(W_1 \cap W_2)$, the probability that both events occur.

- There are 10 socks in the drawer and only 6 are white. The probability that the first sock selected is white is $P(W_1) = \frac{6}{10}$.
- After selecting one white sock, 9 socks remain and 5 of them are white. The probability that the second sock is white given that the first sock was white is $P(W_2|W_1) = \frac{5}{9}$.

Now apply the multiplicative rule:

$$\begin{aligned} P(W_1 \cap W_2) &= P(W_1) \cdot P(W_2 \mid W_1) \\ &= \frac{6}{10} \cdot \frac{5}{9} \\ &= \frac{30}{90} \\ &= \frac{1}{3} \end{aligned}$$

The probability of selecting two white socks is $\frac{1}{3} \approx 0.3333$ or about 33.33%.

Example 6.3.2

A standard deck of cards contains 52 cards, of which 13 are hearts. You draw two cards from the deck without replacement. What is the probability that both cards drawn are hearts?

Solution

Define the events:

H_1 is the first card is a heart and H_2 is the second card is a heart.

We want to find $P(H_1 \cap H_2)$, the probability that both events occur.

Because the cards are drawn without replacement, the second draw depends on the first. This makes the events dependent, so we use the multiplicative rule for dependent events:

$$P(H_1 \cap H_2) = P(H_1) \cdot P(H_2 \mid H_1)$$

On the first draw, the probability of drawing a heart is $P(H_1) = \frac{13}{52}$.

After drawing one heart, 51 cards remain, 12 of which are hearts. S, $P(H_2|H_1) = \frac{12}{51}$.

Now apply the multiplicative rule:

$$\begin{aligned} P(H_1 \cap H_2) &= P(H_1) \cdot P(H_2 \mid H_1) \\ &= \frac{13}{52} \cdot \frac{12}{51} \\ &= \frac{156}{2652} \\ &= \frac{1}{17} \end{aligned}$$

The probability of selecting two hearts is $\frac{1}{17} \approx 0.0588$ or about 5.88%.

You Try It $6.3.A$

A small bag contains 5 candies, 3 of which are strawberry-flavored. You randomly select two candies from the bag without replacement.

What is the probability that both candies selected are strawberry?

Answer available at the end of this section.

Example $6.3.3$

A science club has 10 members, of which 4 are seniors and 6 are juniors. The club president randomly selects two different members to help set up an experiment.

What is the probability that the first member selected is a junior and the second member selected is a senior?

Solution

Define the events:

J_1 is the first member is a junior and S_2 is the second member is a senior.

We want to find $P(J_1 \cap S_2)$.

Because the selections are made without replacement, these events are dependent. Use the multiplicative rule for dependent events:

$$P(J_1 \cap S_2) = P(J_1) \cdot P(S_2 \mid J_1)$$

For the first selection, 6 of the 10 members are juniors so $P(J_1) = \frac{6}{10}$.

After selecting a junior, there are 4 seniors and 9 members remaining: $P(S_2|J_1) = \frac{4}{9}$.

Now apply the multiplicative rule:

$$P(J_1 \cap S_2) = \frac{6}{10} \cdot \frac{4}{9} = \frac{24}{90} = \frac{4}{15}$$

The probability of selecting a junior and then a senior is $\frac{4}{15} \approx 0.2667$ or about 26.67%.

You Try It $6.3.B$

A school library has 15 books on a display shelf: 9 are fiction and 6 are nonfiction. The librarian randomly selects two different books to feature in the morning announcements.

What is the probability that the first book selected is fiction and the second book selected is nonfiction?

Answer available at the end of this section.

Example $6.3.4$

A clinic tracks patient symptoms during flu season. They estimate that there is a 45% chance that a patient will arrive with a fever. If a patient *does* arrive with a fever, there is a 60% chance that the same patient will also have a cough.

What is the probability that a randomly selected patient arrives with both a fever and a cough?

Solution

Define the events:

F is arrives with a fever and C is arrives with a cough.

We want to find $P(F \cap C\backslash)$, the probability that a patient arrives with both a fever and a cough.

Because the probability of having a cough depends on whether the patient has a fever, these events are dependent. Use the multiplicative rule for dependent events:

$$P(F \cap C) = P(F) \cdot P(C \mid F)$$

In the problem we are told $P(F) = 0.45$ and \P(C \mid F)=0.60\).

Multiply:

$$0.45 \cdot 0.60 = 0.27$$

The probability that a patient arrives with both a fever and a cough is 0.27, or 27%.

You Try It $6.3.C$

A delivery company tracks issues with packages during transit. They estimate that there is a 50% chance that a package will arrive late. If a package does arrive late, there is a 40% chance that it will also be damaged.

What is the probability that a randomly selected package arrives both late and damaged?

Answer available at the end of this section.

Probability of More than Two Dependent Events

The multiplicative rule provides a framework for finding the probability that two events occur together, but the same reasoning applies to three or more events as well. The key idea remains the same: we treat the occurrence of each event as a step in a sequence, multiplying by the probability of the next event given that all previous events have already occurred. In other words, intersections such as $P(A \cap B \cap C)$ can be built by successively applying conditional probabilities. This leads to a general form of the rule:

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|(A \cap B).$$

The multiplicative rule naturally scales to any number of events, allowing us to analyze multi-stage processes, repeated selections, and complex dependent situations.

Example 6.3.5

A box contains 6 tiles numbered 1 through 6. You draw three tiles one after another without replacement. What is the probability that all three tiles drawn are even?

Solution

Define the events:

E_1 is first tile is even, E_2 is second tile is even, and E_3 is third tile is even.

We want to find $P(E_1 \cap E_2 \cap E_3)$, the probability that all three tiles are even.

Because tiles are drawn without replacement, each event changes the sample space for the next. These events are dependent, so we use the extended multiplicative rule:

$$P(E_1 \cap E_2 \cap E_3 = P(E_1) \cdot P(E_2 \mid E_1) \cdot P(E_3 \mid (E_1 \cap E_2)$$

- At the start, 3 of the 6 tiles are even so $P(E_1) = \frac{3}{6}$.
- After drawing one even tile, 2 even tiles remain out of 5 total tiles: $P(E_2 \mid E_1) = \frac{2}{5}$.
- After drawing two even tiles, 1 even tile remains out of 4 total tiles:$P(E_3 \mid (E_1 \cap E_2)) = \frac{1}{4}$.

Multiply:

$$\frac{3}{6} \cdot \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{20}$$

The probability that all three tiles drawn are even is $\frac{1}{20} = 0.05$ or 5%.

You Try It 6.3.D

A standard deck has 52 cards, of which 26 are red (hearts and diamonds). You draw three cards from the deck without replacement. What is the probability that all three cards drawn are red?

Answer available at the end of this section.

Independent Events

All examples in the section so far have focused on dependent events, where the outcome of one event changes the probability of the next. This is why we used conditional probabilities such as $P(B|A)$ because once event A happens, the sample space for event B is no longer the same.

However, not all real-world situations work this way. Many processes involve events that do not influence each other, meaning the probability of the second event stays the same no matter what happened first. These are called **independent events.** For independent events, the multiplicative rule still applies, but the rule is simpler.

Recall that when events are independent, $P(B|A) = P(B)$. Substituting into the general rule gives a simplified multiplicative rule for independent events:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B \mid A) \\ &= P(A) \cdot P(B) \end{aligned}$$

Multiplicative Rule for Independent Events

If events A and B are independent, then $P(A \cap B) = P(A) \cdot P(B)$.

The key difference is that dependent events require conditional probabilities, while independent events do not, because the sample space does not change from one event to the next. This shift allows us to analyze situations like repeated coin flips, multiple dice rolls, or any process where outcomes remain unaffected by previous results.

Example 6.3.6

A coin is tossed and a die is rolled. What is the probability that the coin shows heads and the die shows a number less than 3?

Solution

Because the coin toss does not affect the die roll, these events are independent.

Define the events:

A is the event that the coin lands on heads and B is the event that the lands on a number less than 3.

We want to find $P(A \cap B$, the probability that both events occur.

Compute each probability:

- The probability of heads on the coin is $P(A) = \frac{1}{2}$.
- The probability of a number less than 3 on the die is $P(B) = \frac{2}{6}$.

Now apply the multiplicative rule for independent events:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ &= \frac{1}{2} \cdot \frac{2}{6} \\ &= \frac{2}{12} \\ &= \frac{1}{6} \end{aligned}$$

The probability the coin shows heads and the die shows a number less than 3 is $\frac{1}{6} \approx 0.1667$ or 16.67%.

Example 6.3.7

In a drawer, you have 10 pairs of socks, 6 of which are white. You also have 7 t-shirts, 3 of which are white. You randomly reach in and pull out one pair of socks and one t-shirt. What is the probability that both items are white?

Solution

Because choosing a pair of socks does not affect the color distribution of the t-shirts, these events are independent.

Define the events:

A is socks are white and B is t-shirt is white.

We want to find $P(A \cap B)$, the probability that both events occur.

Compute each probability:

- The probability of a pair white socks is $P(A) = \frac{6}{10}$.
- The probability a white t-shirt is $P(B) = \frac{3}{7}$.

Now apply the multiplicative rule for independent events:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ &= \frac{6}{10} \cdot \frac{3}{7} \\ &= \frac{18}{70} \\ &= \frac{9}{35} \end{aligned}$$

The probability that both items are white is $\frac{9}{35} \approx 0.2571$ or 25.71%.

Exercise 6.3.E

A snack shelf contains 8 granola bars, 5 of which are chocolate. A cooler next to it contains 6 bottles of juice, 2 of which are apple. You randomly pick one granola bar and one juice bottle.

What is the probability that both items are chocolate and apple, respectively?

Answer available at the end of this section.

Example 6.3.8

Approximately 9% of smartphones have a manufacturing defect that causes the screen to flicker intermittently. If 3 smartphones are selected at random for inspection, what is the probability that all three have this defect?

Solution

Because one smartphone's condition does not affect the others, these events are independent.

Define the events:

D_1 is first phone has the defect, D_1 is second phone has the defect,and D_1 is third phone has the defect.

We want to find $P(D_1 \cap D_2 \cap D_3)$, the probability that all three phones have the defect.

$$P(D_1 \cap D_2 \cap D_3) = P(D_1) \cdot P(D_2) \cdot P(D_3)$$

In the problem we are told that for any cell phone $P(D) = 0.09$.

Multiply:

$$0.09 \cdot 0.09 \cdot 0.09 = (0.09)^3 = 0.0007$$

The probability that all three cell phones have the defect is 0.0007 or 0.07%.

You Try It $6.3.F$

Carlos is practicing free-throws at basketball practice. The probability that he successfully makes a free-throw is 60%. He will attempt 4 free-throws.

What is the probability that Carlos makes all four free-throws?

Answer available at the end of this section.

Example $6.3.9$

A child is tossing a beanbag onto a small mat. Each time she tosses it, the probability she lands it on the mat is 40%. She will make three tosses.

What is the probability that she lands the first toss, misses the second, and lands the third?

Solution

Because each toss is independent, we can multiply the probabilities of the three events.

The probability of landing the toss is $P(\text{lands}) = 0.40$

The probability of missing the toss is $P(\text{misses}) = 1 - 0.40 = 0.60$.

Multiply for the specific pattern "lands and misses and lands":

$$\begin{aligned} P(\text{lands and misses and lands}) &= P(\text{lands}) \cdot P(\text{misses}) \cdot P(\text{lands}) \\ &= 0.40 \cdot 0.60 \cdot 0.40 \\ &= (0.40)^2 \cdot 0.60 \\ &= 0.096 \end{aligned}$$

The probability she lands the first toss, misses the second, and lands the third is 0.096 or 9.6%.

You Try It $6.3.G$

A spinner has sections of different colors. The probability of landing on a green section is 45%. A player spins the spinner three times.

What is the probability that the spinner lands on green on the first spin, but not green on the second or third spins?

Answer available at the end of this section.

Using the Complement Rule with the Multiplicative Rules

The multiplicative rules work well when we want the probability of a specific sequence of events - such as a basketball player making four shots in a row or a student answering three questions correctly in succession. In these cases, we simply multiply the probability of success each time.

But many practical questions aren't about one exact sequence. Instead, they broader events such as "making at least one shot." Calculating these directly can require listing several different sequences, each with its own probability. That becomes inefficient very quickly.

This is where the *Complement Rule* becomes a powerful shortcut. Rather than computing every possible way an event *can* happen, we often get the answer more easily by finding the probability of what doesn't happen and subtracting from 1.

For example, the probability of "making at least one shot" can be found by computing the probability of "making no shots at all" and taking the complement. Likewise, the probability of "at least one defective item in a batch" is often easier to find by first calculating the probability that "*all items are good*."

Example 6.3.10

Mary is in the archery club. The probability that she hits the target on any attempt is 60%. She will make four attempts. Find the probability that Mary hits the target at least once.

Solution

To find the probability of at least one hit, it is much easier to use the *Complement Rule*. Instead of calculating all the different ways Mary could hit the target at least once, compute the probability of the opposite event: no hits. Then subtract from 1.

To find the probability that Mary gets no hits, calculate the probability of missing all four times.

The probability of a miss is $P(\text{miss}) = 1 - 0.60 = 0.40$.

The probability of four misses in a row is

$$\begin{aligned} &P(\text{misses and misses and misses and misses}) \\ &= P(\text{misses}) \cdot P(\text{misses}) \cdot P(\text{misses}) \cdot P(\text{misses}) \\ &= 0.40 \cdot 0.40 \cdot 0.40 \cdot 0.40 \\ &= (0.40)^4 \\ &= 0.0256 \end{aligned}$$

Now apply the Complement Rule:

$$\begin{aligned} P(\text{at least one hit}) &= 1 - P(\text{miss all } 4) \\ &= 1 - 0.0256 \\ &= 0.9744 \end{aligned}$$

Mary hits the target at least once with probability 0.9744 or 97.44%.

You Try It 6.3.H

A student is trying to log in to an online homework system. Each time she enters her password, the probability that she enters it correctly is 70%. She will make three attempts.

What is the probability that she enters her password correctly at least once?

Answer available at the end of this section.

Section Summary

The *Multiplicative Rules of Probability* give the probability that two events occur together.

For dependent events A and B,

$$P(A \cap B) = P(A) \cdot P(B|A).$$

When event B is independent of event A, the occurrence of A has no effect on the probability of B. For independent events A and B, the rule simplifies to

$$P(A \cap B) = P(A) \cdot P(B).$$

A closely related idea is finding the probability that *at least one* of several events occurs. Instead of adding many overlapping probabilities directly, we use the *Complement Rule*:

$$P(\text{at least one}) = 1 - P(\text{none}).$$

To compute $P(\text{none})$, apply the *Multiplicative Rule* to the sequence of "non-occurrence" events.

Section 6.3 Review Questions

1. A gym reports that 53% of its members participate in the strength-training program. Furthermore 27% of the strength-training participants take part in the cycling class. If one member is chosen at random, determine the probability that the member participates in both strength training and cycling.

2. At a community movie night, 90% of attendees buy popcorn. It is also known that 65% of those who buy popcorn purchase a soft drink. If one attendee is selected at random, determine the probability that the attendee buys both popcorn and a soft drink.

3. The probability that Sam parks in a no-parking zone at school is 20%. The probability that Sam gets a ticket when he parks in a no-parking zone is 30%. On Tuesday, Sam arrives at school. Find the probability that Sam parks in a no-parking zone and gets a parking ticket on Tuesday.

4. A science classroom keeps 30 lab samples, and 5 of them are mineral specimens. If 2 samples are chosen at random without replacement, determine the probability that both selected samples are mineral specimens..

5. A flashlight has 6 batteries, 2 of which are defective. If 2 are selected at random without replacement, find the probability that both are defective.

6. A wildlife center is caring for 8 rescued turtles, and 5 of them are currently undergoing medical treatment. If 3 turtles are selected at random without replacement, determine the probability that all three selected turtles are receiving treatment.

7. A jar with M&M's has 3 red, 2 blue and 5 brown. One candy is selected and then a second candy is selected without replacing the first one. Find each probability.
   a. both candies are blue
   b. the first candy is blue and then second one is brown
   c. no red candies are selected
   d. at least one red candy is selected.

8. Say if you suspect events *A* and *B* in each scenario are *independent* or *not independent*.
   a. A fair coin is tossed twice.
      i. Event *A*: First toss is a head.
      ii. Event *B*: Second toss is a head.
   b. Weather is observed tomorrow in Washington, DC and Baltimore.
      i. Event *A*: It will rain tomorrow in Washington, DC.
      ii. Event *B*: It will rain tomorrow in Baltimore.
   c. You draw a card from a deck, then draw another card without replacing the first one.

   i. Event *A*: First card is a red card.
   ii. Event *B*: Second card is a red card.
d. You draw a card from a deck, replace the card, and then draw a second card.
   i. Event *A*: First card is a red card.
   ii. Event *B*: Second card is a red card.

9. At a large music festival, 30% of attendees are wearing band merchandise. If 4 attendees are selected at random, assuming each selection is independent, determine the probability that all four are wearing band merchandise.

10. A city's public library reports that 80% of its patrons have an active library card. If 3 patrons are selected at random, determine the probability that
    a. all three have an active library card.
    b. none of the three have an active library card.
    c. least one of the three has an active library card.

11. A carnival booth has a ring-toss game where players successfully land a ring on the peg 75% of the time. You play the game four times. Find the probability that
    a. you land a ring on the peg every time.
    b. you land a ring only on the second toss.
    c. you land a ring on a peg at least once.

Answers

You Try It $6.3.A$

$\frac{3}{10} = 0.3 = 30\%$

You Try It $6.3.B$

$\frac{9}{35} \approx 0.2571 \approx 25.71\%$

You Try It $6.3.C$

20%

You Try It $6.3.D$

$\frac{2}{17} \approx 0.1176 \approx 11.76\%$

Exercise $6.3.E$

$\frac{5}{24} \approx 0.2083 \approx 20.83\%$

You Try It $6.3.F$

$0.1296 = 12.96\%$

You Try It $6.3.G$

$0.136125 = 13.6125\%$

You Try It $6.3.H$

$0.973 = 97.3\%$

This page titled 6.3: The Multiplicative Rules of Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

6.4: Tree Diagrams and Multistage Probability

Learning Objectives

- Represent multistage probability scenarios with tree diagrams.
- Solve conditional probability problems using tree diagrams.
- Solve reverse conditional probability problems using an intuitive Bayes' Theorem approach.

Introduction

In many probability settings, especially those involving **multiple stages**, it becomes helpful to visualize how outcomes unfold step by step. A **probability tree diagram** provides this structure by displaying each stage as a branching path, with every branch labeled by its corresponding probability. Tree diagrams make complex, multistage experiments easier to interpret, compare, and calculate. As you move from left to right along the tree, you can trace complete outcome sequences and compute their probabilities by multiplying along the branches.

In this section, you will explore probabilities of more complex events by using tree diagrams as a tool for problem solving

Drawing a Probability Tree

In multistage probability experiments, listing every possible outcome in a sample space can quickly become inefficient -- especially when the outcome of one stage determines what happens in the next. A **probability tree diagram** provides a clear and systematic way to represent all possible outcomes and their associated probabilities. By displaying each stage as a branching structure, a tree diagram allows us to trace complete paths through the experiment and compute their probabilities using the *Multiplicative Rule of Probability*.

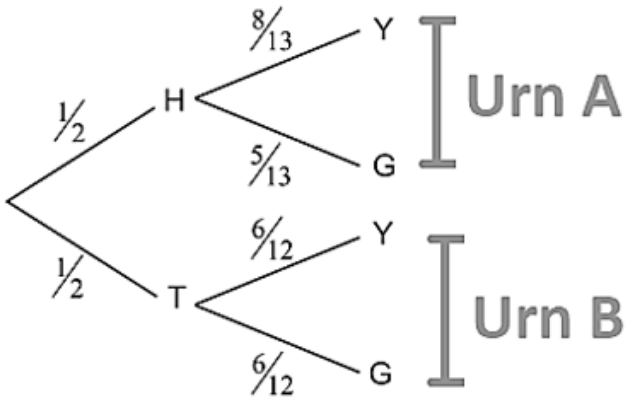
Consider the following two-stage game of chance: First, a coin is flipped. Then, depending on the coin's result, a chip is drawn from one of two urns.

- If the coin lands heads, a chip is drawn from Urn A, which contains 8 yellow chips and 5 green chips.
- If the coin lands tails, a chip is drawn from Urn B, which contains 6 yellow chips and 6 green chips.

Because the chip selection depends on the coin toss, there are several possible ways to obtain a yellow or green chip. Rather than listing all outcomes, we construct a probability tree diagram to organize the information using the steps below.

1. **Identify the stages.** In this game the first stage is the coin toss and the second stage is the chip selection from the appropriate urn.
2. **Draw a set of branches for the first stage.** From a starting point, draw two branches representing the coin outcomes - one for heads and one for tails. Label each branch with its probability:
    - $P(H) = \frac{1}{2}$
    - $P(T) = \frac{1}{2}$
3. **Draw sets of branches for the second stage.** You will need two sets of branches, one for drawing a chip if the coin was Heads and one for drawing a chip if the coin was Tails.
    - From the Heads branch, draw branches for drawing a chip from Urn A and label each with its probability.
        - $P(Y) = \frac{8}{13}$
        - $P(G) = \frac{5}{13}$
    - From the Tails branch, draw branches for drawing a chip from Urn B and label each with its probability.
        - $P(Y) = \frac{6}{12}$
        - $P(G) = \frac{6}{12}$

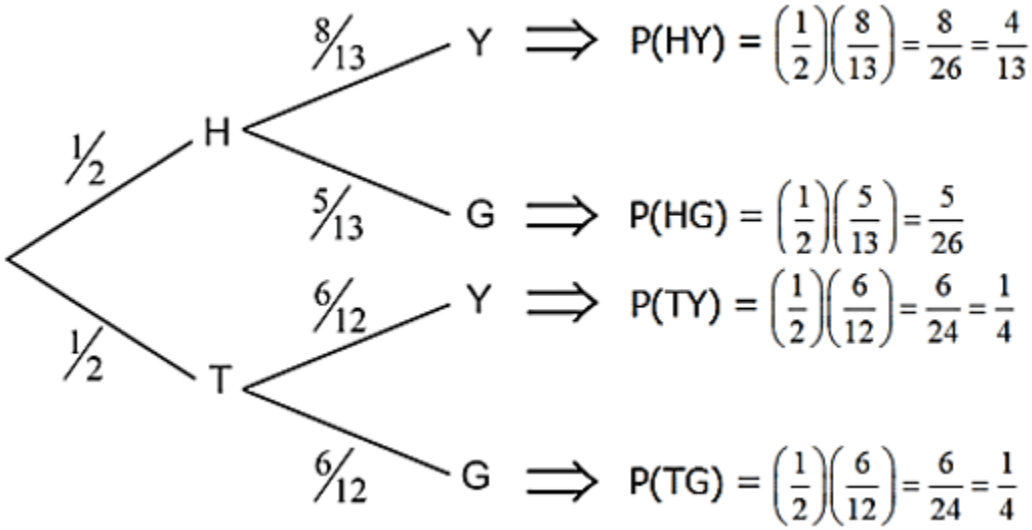
So far, the tree diagram should look like this:



4. **Compute the probability of each complete outcome.** Apply the *Multiplicative Rule of Probability* by multiplying across each path.

   - Heads → Yellow from Urn A: $\frac{1}{2} \cdot \frac{8}{13} = \frac{8}{26} = \frac{4}{13}$
   - Heads → Green from Urn A: $\frac{1}{2} \cdot \frac{5}{13} = \frac{5}{26}$
   - Tails → Yellow from Urn B: $\frac{1}{2} \cdot \frac{6}{12} = \frac{6}{24} = \frac{1}{4}$
   - Tails → Green from Urn B: $\frac{1}{2} \cdot \frac{6}{12} = \frac{6}{24} = \frac{1}{4}$

The completed probability tree is shown below:



5. **Check that the probabilities of all endpoints sum to 1.** The endpoints of the tree represent the entire sample space. Their probabilities must add to 1, confirming that the diagram is complete and accurate.

$$\frac{4}{13} + \frac{5}{26} + \frac{1}{4} + \frac{1}{4} = 1$$

The tree diagram is now ready to be used to answer questions about various probabilities as shown in Example 6.4.1.

Using a Probability Tree to Solve Problems

Example 6.4.1

In the game of chance described previously, you win if you end up with a yellow chip.

What is the probability of getting a yellow chip?

Solution

Looking at the diagram, there are two ways to select a yellow chip:

1. Get Heads on the coin and yellow from Urn A:

$$P(\text{Heads} \cap \text{Yellow}) = \tfrac{4}{13}$$

2. Get Tails on the coin and yellow from Urn B:

$$P(\text{Tails} \cap \text{Yellow}) = \tfrac{1}{4}$$

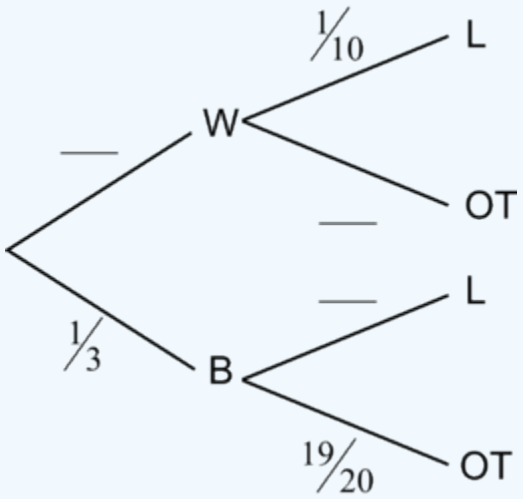
Add the probabilities of these "two paths" to getting a yellow chip:

$$\begin{aligned} P(\text{yellow}) &= P(\text{Heads} \cap \text{Yellow}) + P(\text{Tails} \cap \text{Yellow}) \\ &= \frac{4}{13} + \frac{1}{4} \\ &= \frac{16}{52} + \frac{13}{52} \\ &= \frac{29}{52} \end{aligned}$$

The probability of getting a yellow chip is $\frac{29}{52} \approx 0.5577$ or 55.77%.

Example 6.4.2

Sarah either walks (W) or rides the bus (B) to school. Her likelihood of being late (L) or on time (OT) depends on which mode of transportation she chooses. The partial tree diagram displays these stages in sequence.



Use the probabilities given in the tree diagram to complete the tree diagram and find the probability that Sarah is late for school.

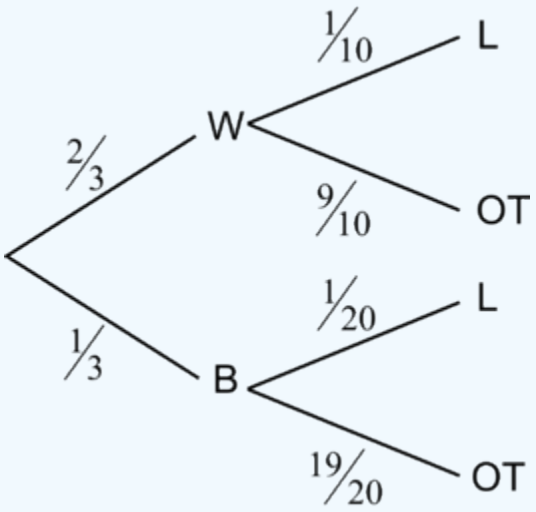
Solution

First, recall that the events on each set of branches are complementary events. That is, they must add to be 1. This allows you to find the all the missing probabilities in the tree:

$P(W) = 1 - P(B) = 1 - \frac{1}{3} = \frac{2}{3}$

$P(OT|W) = 1 - P(L|W) = 1 - \frac{1}{10} = \frac{9}{10}$

$P(W|B) = 1 - P(OT|B) = 1 - \frac{19}{20} = \frac{1}{20}$



Now the diagram is complete, each full path represents one possible outcome of Sarah's trip to school, and the probability of that outcome is found by multiplying the probabilities along its branches. To determine the probability that Sarah is late, we identify the two distinct paths in the tree that end in L:

1. Walk and Late: $P(W \cap L) = \frac{2}{3} \cdot \frac{1}{10} = \frac{2}{30}$
2. Bus and be Late: $P(B \cap L) = \frac{1}{3} \cdot \frac{1}{20} = \frac{1}{60}$

$$\begin{aligned} P(L) &= P(W \cap L) + P(B \cap L) \\ &= \frac{2}{30} + \frac{1}{60} \\ &= \frac{4}{60} + \frac{1}{60} \\ &= \frac{5}{60} \\ &= \frac{1}{12} \end{aligned}$$

The probability that Sarah is late is $\frac{1}{12} \approx 0.0833$ or 8.33%.

Example 6.4.3

An assembler of computer routers and modems purchases its parts from two suppliers.

- Company A provides 60% of the parts.

- Company B provides the remaining 40%.

Quality records show that 3% of Company A's parts are defective, while 8% of Company B's parts are defective. If one part is selected at random, determine the probability that it is defective.

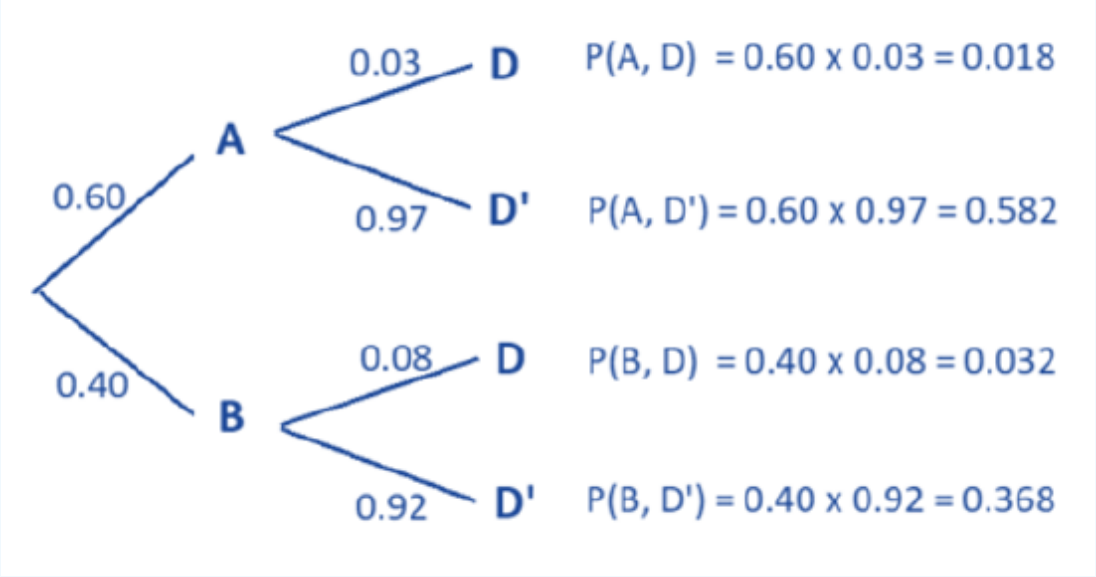
Solution

To organize this two-stage situation, we use a probability tree diagram. The first set of branches represents the suppliers (A) or (B). These probabilities are $P(A) = 0.60$ and $P(B) = 0.40$.

The second set of branches represents whether the selected part is defective ((D\) or not defective D'. Notice, however, that these probabilities change depending on which company made the part.

- If Company A makes the part, $P(D|A) = 0.03$ and $P(D'|A) = 1 - 0.03 = 0.97$.
- If Company B makes the part, $P(D|B) = 0.08$ and $P(D'|B) = 1 - 0.08 = 0.92$.

We can determine the probabilities of the multistage events by applying the *Multiplicative Rule of Probability* and multiplying through the branches:



In the completed tree diagram, verify the sum of the probabilities of all outcomes add to be 1:

$$0.018 + 0.582 + 0.032 + 0.368 = 1$$

To determine the probability an item is defective, we identify the two distinct paths in the true where the item is defective:

1. A and Defective: $P(A \cap D) = 0.60 \cdot 0.03 = 0.018$
2. B and Defective: $P(B \cap D) = 0.40 \cdot 0.08 = 0.032$

$$\begin{aligned} P((\text{item is defective}) &= P(A \cap D) + P(B \cap D) \\ &= 0.018 + 0.032 \\ &= 0.05 \end{aligned}$$

The probability that an item is defective 0.05 or 5%.

You Try It 6.4.A

A wildlife researcher is studying two species of birds in a forest where 40% of the birds in the area are sparrows (S) and the remaining 60% are finches (F).

During observation, the researcher notes that 15% of sparrows and 25% of finches sing during the morning observation period.

If one bird is spotted at random, what is the probability that it sings during the morning observation period?

Answer available at the end of this section.

Example 6.4.4

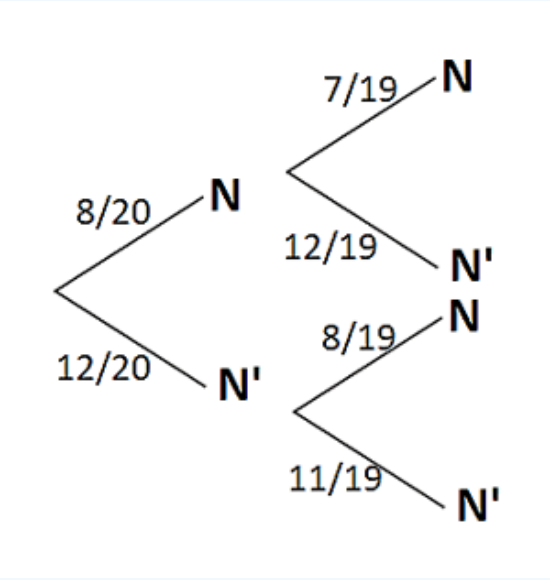
A box contains 20 candies, of which 8 contain nuts and 12 do not. Two candies are selected at random without replacement. Determine the probability that exactly one of the two selected candies contains nuts.

Solution

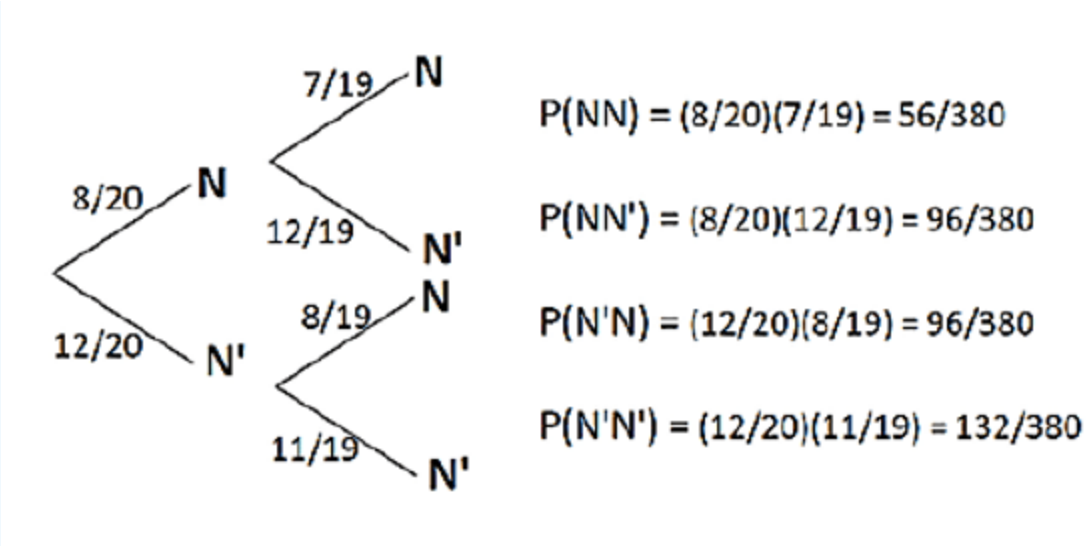
We proceed as we did in Example 6.4.1 by first drawing a tree diagram and labeling each branch with the probability of getting a piece of candy with nuts N and the probability of a piece without nuts N'. For the first selection, these probabilities are $P(N) = \frac{8}{20}$ and $P(N') = \frac{12}{20}$.

Notice, however, that on the second selection, the probabilities change depending on the first selection.

- If the first candy has nuts, there are 7 pieces remaining with nuts and 12 pieces remaining without nuts. There are 19 pieces in all. $P(N_2|N_1) = \frac{7}{19}$ and $P(N_2'|N_1) = \frac{12}{19}$.
- If the first candy doesn't have nuts, there are 8 pieces remaining with nuts and 11 pieces remaining without nuts. There are 19 pieces in all. $P(N_2|N_1') = \frac{8}{19}$ and $P(N_2'|N_1') = \frac{11}{19}$.



We can determine the probabilities of the multistage events by applying the *Multiplicative Rule of Probability* and multiplying through the branches:



In the completed tree diagram, verify the sum of the probabilities of all outcomes add to be 1:

$$\frac{56}{380} + \frac{96}{380} + \frac{96}{380} + \frac{132}{380} = \frac{380}{380} = 1$$

To determine the probability of getting exactly one piece of candy with nuts, we identify the two distinct paths in the tree:

1. Nuts and No nuts: $P(N_1 \cap N_2') = \frac{8}{20} \cdot \frac{12}{19} = \frac{96}{380}$
2. No nuts and Nuts: $P(N_1' \cap N_2) = \frac{12}{20} \cdot \frac{8}{19} = \frac{96}{380}$

$$\begin{aligned} P((\text{exactly one piece with nuts}) &= P(N_1 \cap N_2') + P(P(N_1' \cap N_2)) \\ &= \frac{96}{380} + \frac{96}{380} \\ &= \frac{192}{380} \\ &= \frac{48}{95} \end{aligned}$$

The probability of getting exactly one piece of candy with nuts is $\frac{48}{95} \approx 0.5053$ or 50.53%.

You Try It $6.4.B$

A jar contains 10 marbles, of which 4 are red and 6 are blue. Two marbles are selected at random without replacement. Determine the probability that exactly one of the two selected marbles is red.

Answer available at the end of this section.

Reversing the Condition: An Intuitive Approach to Bayes' Theorem

So far, we have worked with conditional probability, dependent events, and probability trees. In many of those situations, we knew the chance that one event happened first, and then the chance of a second event happening after it. For example, we might know the probability that a student takes the bus and the chance that a student is late if they take the bus.

However, many everyday problems work in the *reverse* direction. We observe a result first and then we want to reason backward to figure out the most likely cause. That is, if we know a student is late, what's the chance they took the bus? This “reverse” conditional probability is extremely common in real life.

This type of backwards reasoning is exactly what a famous rule in probability, *Bayes' Theorem*, is designed to handle. ***Bayes' Theorem*** helps us revise a probability after new information is observed. Instead of asking, "What is the chance of this result given the cause?" we ask, "What is the chance of the cause given the result?"

To illustrate this idea, recall the scenario from Example 6.4.3. An assembler of computer routers and modems purchases parts from two suppliers. Company A provides 60% of the parts and Company B provides the remaining 40%. It turns out that 3% of Company A's parts are defective while 8% of Company B's parts are defective.

From this information, we can easily answer a "forward" question such as:

What is the probability a part is defective given it came from Company A? This is simply 3%.

But now consider the reverse question:

If the part I just bought is defective, what is the probability it came from Company A? This is no longer 3%.

This is a very different question—and it is exactly the kind of question Bayes' Theorem helps us answer.

Although Bayes' Theorem has a formal algebraic formula, using that formula directly is beyond the scope of this course. Fortunately, Bayes' Theorem problems are very accessible when we use probability trees. A tree diagram organizes the information step-by-step and allows us to compute reverse conditional probabilities by comparing the correct branches. In other words, we do not need the formula to understand the idea. We only need a clear tree diagram and careful reasoning.

Let's take a look at an example.

Example 6.4.5

From Example 6.4.3:

An assembler of computer routers and modems purchases its parts from two suppliers.

- Company A provides 60% of the parts.
- Company B provides the remaining 40%.

Quality records show that 3% of Company A's parts are defective, while 8% of Company B's parts are defective.

A part was just found to be defective. What's the chance the part was provided by Company A?

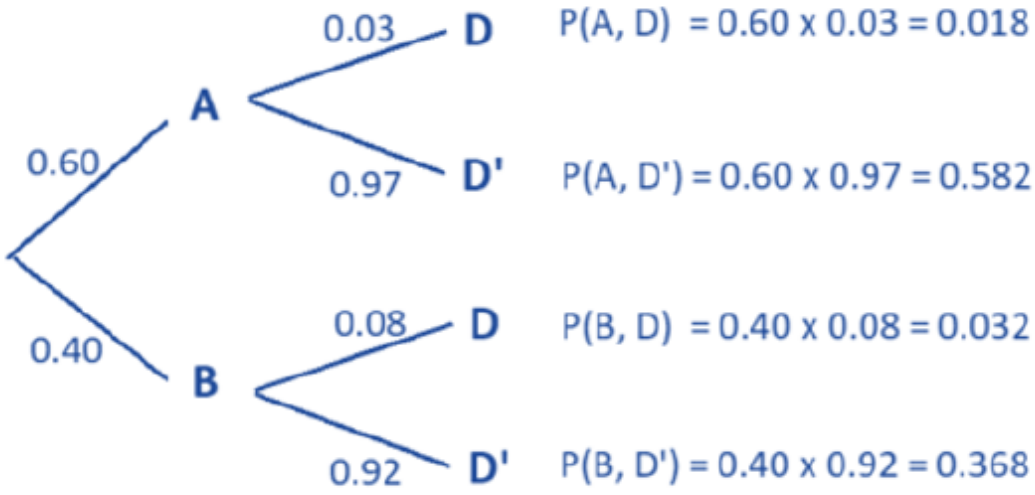
Solution

We want to find the probability that a defective part came from Company A. This is a *reverse* conditional probability: instead of asking for the chance a part is defective given it came from A, we ask for the chance it came from A given that it is defective. That is, we want to calculate $P(A|D)$. Probability trees make this type of question clear and manageable.

To compute $P(A|D)$ we need two quantities:

- The probability a part came from Company A and is defective: $P(A \cap D)$
- The overall probability that a part is defective: $P(D)$

In the solution of Example 6.4.3 we drew this probability tree:



From the tree diagram we see that

- $P(A \cap D) = 0.60 \cdot 0.03 = 0.018$
- $P(D) = 0.018 + 0.032 = 0.05$

Substituting into the reversed conditional formula,

$$\begin{aligned} P(A \mid D) &= \frac{P(A \cap D)}{P(D)} \\ &= \frac{0.018}{0.05} \\ &= 0.36 \end{aligned}$$

There is a 36% chance that the defective part was provided by Company A.

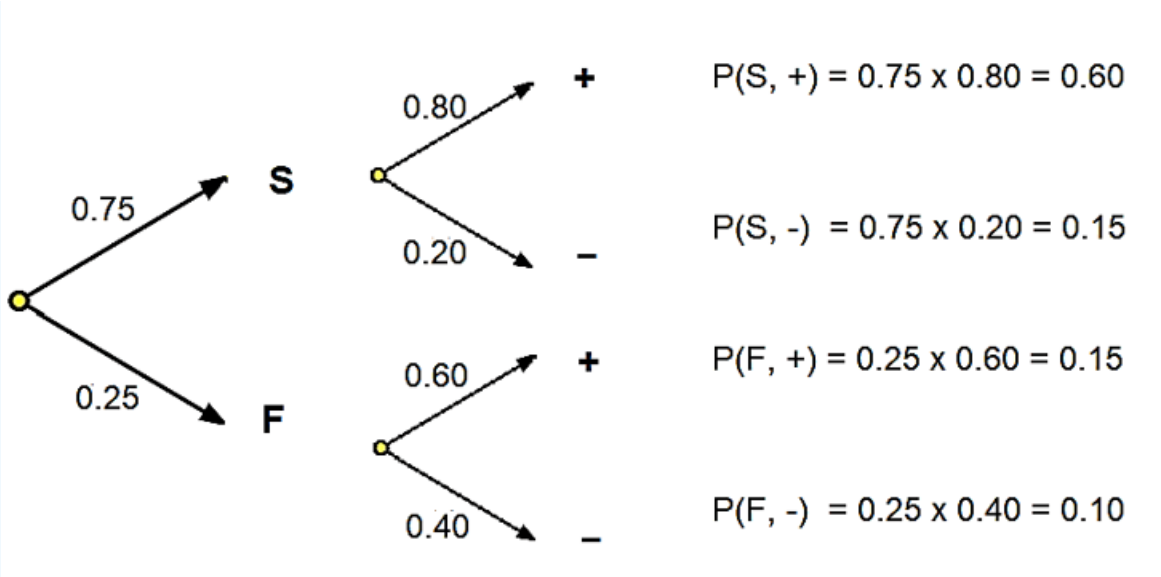
Example 6.4.6

A restaurant offers two entrée choices: steak or fish. About 75% of customers order steak, and the remaining 25% order fish. After eating, each customer is asked whether they were satisfied with their meal. A survey shows: 80% of steak-ordering customers are satisfied and 60% of fish-ordering customers are satisfied.

a. Construct a probability tree for this scenario.

Solution

Let S be ordered steak, F be ordered fish, $+$ be satisfied, and $-$ be not satisfied.



b. Find the probability a customer is satisfied.

Solution:

A customer is satisfied in two cases:

1. They ordered steak and were satisfied: $P(S, +) = 0.75 \cdot 0.80 = 0.60$
2. They ordered fish and were satisfied: $P(F, +) = 0.25 \cdot 0.60 = 0.15$

$$P(+) = P(S, +) + P(F, +) = 0.60 + 0.15 = 0.75$$

There is a 75% chance a randomly selected customer is satisfied.

c. Find the probability a customer ordered fish given they were satisfied

Solution

Now we restrict the sample space to satisfied customers only: (S,+) and (F,+). This probability was found in part b.

Use the conditional probability formula:

$$\begin{aligned} P(S \mid +) &= \frac{P(S \cap +)}{P(+)} \\ &= \frac{0.60}{0.75} \\ &= 0.68 \end{aligned}$$

If a customer was satisfied, there is an 80% chance they ordered steak.

Example 6.4.7

Using the same information and probability tree as in Exercise 6.4.6, find the probability a customer ordered fish given they were dissatisfied.

Solution

First restrict the sample space to those customers who are dissatisfied. There are two cases:

1. They ordered steak and were dissatisfied: $P(S, -) = 0.75 \cdot 0.20 = 0.15$
2. They ordered fish and were dissatisfied: $P(F, -) = 0.25 \cdot 0.40 = 0.10$

$$P(-) = P(S, -) + P(F, -) = 0.15 + 0.10 = 0.25$$

There is a 25% chance a randomly selected customer is dissatisfied.

Use the conditional probability formula to find the chance that a dissatisfied customer ordered fish:

$$\begin{aligned} P(F \mid -) &= \frac{P(F \cap -)}{P(-)} \\ &= \frac{0.10}{0.25} \\ &= 0.40 \end{aligned}$$

If a customer was dissatisfied, there is an 40% chance they ordered fish.

You Try It $6.4.C$

A school nurse is checking for a common seasonal allergy among students: 30% of the students at the school have the allergy and rest don't.

- If a student has the allergy, the chance that they report sneezing during recess is 70%.
- If a student does not have the allergy, the chance that they report sneezing during recess is 20%.

A student reports sneezing during recess. Find the probability that this student has the allergy.

Answer available at the end of this section.

Section Summary

Probability tree diagrams are a powerful way to organize and solve multistage probability problems. Each branch represents a step in the process, and the probability of any complete outcome is found by multiplying along the branches that form that path. This is an application of the *Multiplicative Rule of Probability*, which tells us how to compute the probability of "and" events.

Once the full tree is built, we can answer many different questions by combining the probabilities of the appropriate outcomes in the sample space. Sometimes this means adding several path probabilities together to find the chance of a broader event.

In some problems, we observe a result first and want to work backward to find the most likely cause. These are called reverse conditional probability problems. To solve them, we restrict the sample space to only the outcomes that match the observed result. Bayes' Theorem is the formal rule that describes this process, but in this course, we can solve these problems effectively by carefully analyzing a tree diagram and comparing the relevant branches.

Tree diagrams give us a clear, visual way to understand both forward and reverse conditional probability, making even complex situations easier to reason about.

Section 6.4 Review Questions

1. A box contains 3 white and 5 green balls. Two balls are drawn from the box in succession without replacement. Illustrate the process by drawing a condensed tree diagram. Find the probabilities for all the outcomes. Then, use the tree diagram to answer the questions.
   a. Find the probability that both balls are white.
   b. Find the probability that exactly one ball is white.
   c. Find the probability that no balls are white.
   d. Find the probability that both balls are the same color.

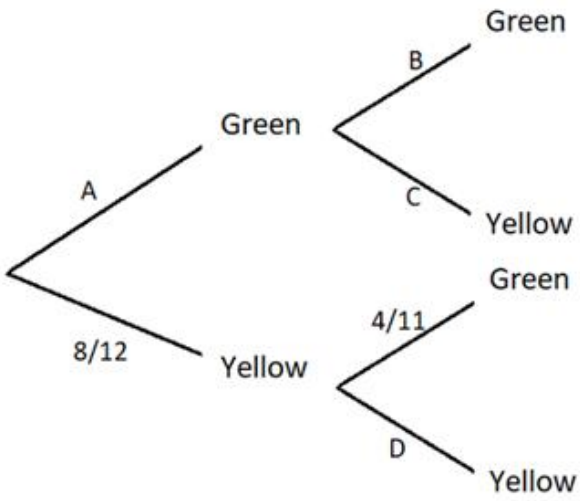
e. Find the probability that the second ball is white.

2. Jamie and Olivia are best friends and neighbors. As such, they often eat dinner together at one of their houses. About 30% of the time, they eat at Jamie's house and her mother makes a non-vegetarian meal 65% of the time. The rest of the time, they eat at Olivia's house and her mother serves a vegetarian meal 55% of the time. Make a tree diagram, including the appropriate probabilities, to illustrate this scenario and then find these probabilities.
   a. What is the probability of going to Olivia's house and eating a non-vegetarian meal?
   b. What is the probability that the girls will eat a vegetarian dinner?
   c. What is the probability the girls went to Olivia's house if you know they ate a vegetarian dinner?

3. A community center offers outdoor yoga classes. The probability that the park is crowded on a given day is 0.6. If the park *is crowded*, the chance that a group of friends decides to attend yoga is 0.2. If the park is *not crowded*, the chance they attend yoga increases to 0.8. Make a tree diagram, including the appropriate probabilities, to illustrate this scenario and then find these probabilities.
   a. Are the events "park is crowded" and "attend yoga" independent?
   b. What is the probability that the park is crowded and the friends attend yoga?
   c. What is the probability that the park is *not* crowded and the friends attend yoga?
   d. What is the probability that the friends attend yoga?
   e. The friends attended yoga today. What is the probability that the park was crowded?

4. It is estimated that 30% of taxpayers "cheat" on their income tax return. Of those who cheat, 60% are audited. Of those who do not cheat, only 20% are audited. Make a tree diagram, listing all final outcomes and the probabilities for those outcomes.
   a. What is the probability that a taxpayer cheats and is also audited?
   b. What is the probability that a taxpayer is audited?
   c. Sue received a letter that she will be audited. What is the probability that Sue cheated on her taxes?

5. A bag contains 12 marbles -- some are green and the rest are yellow. Brian randomly takes a marble from the bag, records its color, and does not replace it. He then repeats this process a second time. He draws the following tree diagram.



a. Give the values of the probabilities labeled *A, B, C* and *D* for the tree diagram.
b. Find the probability that Brian gets two marbles with the same color.
c. Find the probability that Brian gets a yellow marble on the second draw.

6. Ann and Barbara are playing a tennis match. The first player to win 2 sets wins the match. For any given set the probability that Ann wins the set is $\frac{2}{3}$. Illustrate the scenario by drawing a tree diagram to find the probabilities for each possible outcome of the match. (There should be 6 outcomes for the sample space of multistage outcomes.)
   a. Find the probability that Ann wins the match.
   b. Find the probability that 3 sets are played before a winner can be determined.
   c. Find the probability that the player who wins the first set goes on to win the match.

7. Assume that the probability of having skin cancer is 1%. Suppose there is a test for checking the presence of skin cancer. When cancer is present, the test result is positive 90% of the time and negative the other 10%. When cancer is not present, the test is positive 10% of the time, and negative the other 90%.
   a. If someone gets a positive test result, what is the probability they have cancer?
   b. If someone gets a negative test result, what is the probability they do not have cancer?

Answers

You Try It $6.4.A$

0.21 or 21%

You Try It $6.4.B$

$\frac{8}{15} \approx 0.5333 \text{or } 53.33\%$

You Try It $6.4.C$

0.6 = 60%

This page titled 6.4: Tree Diagrams and Multistage Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

6.5: Expectation Involving Chance

Learning Objectives

- Find the expected value in an applied scenario.
- Interpret measures of expectation.

Introduction

In this section, we explore the idea of *expectation* -- a numerical description of what we anticipate when a chance process is repeated many times. Although the outcome of any single trial of a random experiment is unpredictable, the long-run behavior of the experiment follows a stable, predictable pattern. Expected value provides a mathematical way to describe that long-run average.

The key measure that captures this behavior is the **mean** of the possible outcomes in a probability model. We denote this mean by **expected value**, using the symbol μ, the same notation used for the mean in the normal model. In probability, μ represents the value we expect on average over many repetitions of the experiment.

Expected value is one of the most useful concepts in probability. It appears in everyday contexts such as insurance, financial decision-making, and evaluating risks. It also plays a central role in the design of casino games and lotteries - an idea those industries hope most people never fully understand.

Expected Value

To build intuition for the idea of *expected value*, consider a simple game played by Ro and Larson.

Ro rolls a standard six-sided die.

- If the die shows an even number, Ro receives $1 from Larson.
- If it shows an odd number, he gives $1 to Larson.

Either player could win $1 on any roll. The question is: *Who would come out ahead if they played this game many times?*

Because a fair die has three even faces and three odd faces, we expect Ro to win half the time and lose half the time. Over the long run, he should break even. We can confirm this by calculating his average gain or loss:

$$\begin{aligned}
&\text{Ro's expected gain or loss} \\
&= (\text{amount Ro wins} \times P(\text{Ro wins})) + ((\text{amount Ro loses} \times P(\text{Ro loses})) \\
&= \left(\$1 \times \frac{1}{2}\right) + \left(-\$1 \times \frac{1}{2}\right) \\
&= \$0.50 + (-\$0.50) \\
&= \$0
\end{aligned}$$

This value, $\$0$, is Ro's **expected value**. It tells us that if they play the game many times, Ro should neither gain nor lose money overall. A positive expected value would indicate a long-term gain; a negative expected value would indicate a long-term loss.

Importantly, expected value does not predict the outcome of any single roll -- it describes only the long-run average behavior of the game.

This example illustrates the general idea behind expected value: we multiply each possible gain or loss by the probability of that outcome, and then add the products.

$$\mu = \sum_{i=1}^{n} x_i \cdot P(x_i)$$

This theoretical definition captures the same long-run average behavior we observed in Ro and Larson's game and will guide our work with expected value throughout this section.

Expected Value

Expected Value (μ) is the mean outcome if an experiment or procedure with a numerical outcome is repeated many times.

$$\begin{aligned}\mu &= \sum_{i=1}^{n} x_i \cdot P(x_i) \\ &= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + x_3 \cdot P(x_3) + ... + x_n \cdot P(x_n)\end{aligned}$$

where each x is the net amount gained or lost on each outcome of the experiment and $P(x)$ is the probability of that outcome.

Example 6.5.1

Based on attendance records, Ms. Carter has determined the probabilities for the number of students absent from homeroom each day, denoted by x. Find the expected value of x and interpret it.



Solution

The expected value is the the long-run mean outcome. We cannot simply compute $\frac{0+1+2+3+4}{5} = 2$ because each value of x does not occur with equal probability. For instance, having $x = 1$ student absent happens far more often than having $x = 4$ students absent. Each outcome must be weighted by its probability.

$$\begin{aligned}\mu &= \sum_{i=1}^{n} x_i \cdot P(x_i) \\ &= (0 \cdot 0.22) + (1 \cdot 0.34) + (2 \cdot 0.20) + (3 \cdot 0.18) + (4 \cdot 0.06) \\ &= 0 + 0.34 + 0.40 + 0.54 + 0.24 \\ &= 1.52\end{aligned}$$

On average, Ms. Carter can expect 1.52 students to be absent from homeroom each day. Of course, there will never be exactly 1.52 students absent, but over many days the average number of absences should be close to 1.52.

You Try It 6.5.A

The table shows the probabilities that a customer orders x toppings when purchasing a hot dog from a street vendor:



What is the expected number of toppings sold per hot dog?

Answer available at the end of this section.

Example 6.5.3

A university medical research center finds that a certain treatment has a significant success rate in treating a certain type of skin cancer. Suppose five patients receive this treatment. The table shows the probabilities that 1, 2, 3, 4, or all 5 patients are cured.



a. What is the probability that none of the 5 patients are cured?

Solution

Use the fact that the probabilities of all possible outcomes must sum to 1.

$$P(x=0)+0.029+0.132+0.309+0.360+0.168=1$$
$$P(x=0)+0.998=1$$
$$P(x=0)=1-0.998$$
$$P(x=0)=0.002$$

So, the probability that none of the five patients are cured is 0.002.

b. When 5 patients are treated, how many patients can you expect to be cured?

Solution

Compute the expected value using expected value.

$$\mu=\sum_{i=1}^{n}x_i\cdot P(x_i)$$
$$=(0\cdot 0.002)+(1\cdot 0.029)+(2\cdot 0.132)+(3\cdot 0.309)+(4\cdot 0.360)+(5\cdot 0.168)$$
$$=0+0.029+0.264+0.927+1.44+0.84$$
$$=3.5$$

On average, 3.5 patients out of 5 will be cured.

c. What is the mode number of patients who will be cured?

Solution

The mode is the value with the highest probability. Here, $x=4$ has the greatest probability (0.360), so the mode is 4 patients.

d. What is the median number of patients who will be cured?

Solution

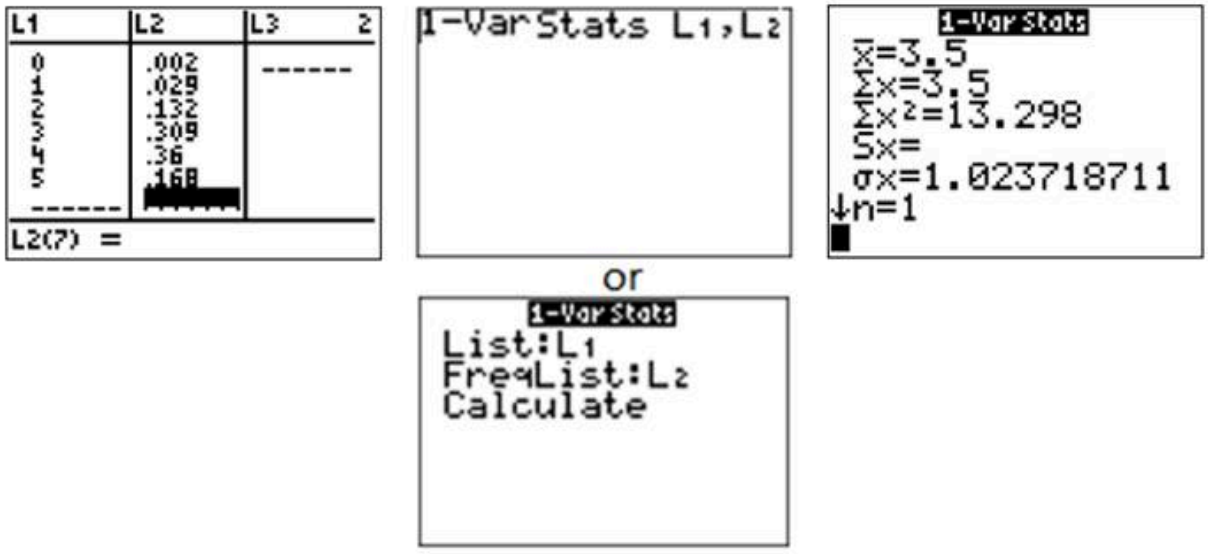
The median is the 50th percentile. Add probabilities until reaching at least 0.50. The cumulative probability exceeds 0.50 when reaching $x=4$, so the median is 4 patients.

Calculation of the expected value can sometimes be tedious, so it is helpful to use technology. Here are the steps for finding the expected value on the calculator for the scenario in Example 6.5.3:

Enter the values of outcomes *x* into L1 and the corresponding probabilities into L2.

Press [STAT] and use the right arrow to move to the CALC menu. Choose 1:1-Var Stats.

Type L1,L2 to indicate you are calculating based on a probability distribution. If you have STAT WIZARD on the newer version of the TI calculator, choose L2 for the FreqList.



Example 6.5.4

A company plans to invest in a particular construction project. There is a 35% chance that it will lose $30,000, a 40% chance that it will break even, and a 25% chance that it will make a profit of $55,000.

How much can the company expect to make or lose on this project?

Solution

It is helpful to make a table that summarizes all the information before finding the expected value. Represent the the loss outcome with a negative number, the break even outcome with 0, and the profit outcome with a positive number.

[redacted]

Now use the formula for the expected value:

$$\begin{aligned}\mu &= \sum_{i=1}^{n} x_i \cdot P(x_i) \\ &= (-30000 \cdot 0.35) + (0 \cdot 0.40) + (55000 \cdot 0.25) \\ &= -10500 + 0 + 13750 \\ &= 3250\end{aligned}$$

The company has an expectation, or average expected gain, of $3,250. If this company makes investments in projects like this one with these particular probabilities and amounts, in the long run it would make an average profit of $3,250 per project. However, on any individual project, the company many gain money, lose money, or break even.

You Try It 6.5.B

The organizers at a balloon festival know that if it's windy in town, only 9,000 people will show up for the festival. But if it's not windy, 30,000 people will show up. The probability of wind is 45%.

What is the expectation for the number of people who attend?

Answer available at the end of this section.

Example 6.5.5

You toss two fair coins simultaneously and record the number of heads. Then you repeat the experiment again and again.

What should be the average number of heads once this experiment has been performed many, many times?

Solution

To answer this, we compute the expected value of the number of heads. Since each coin can land heads or tails, there are four equally likely outcomes in the sample space: $S = \{HH, HT, TH, TT\}$.

We let x be the number of heads in each possible outcome in the sample space.

- For HH, $x = 2$ heads.
- For HT, $x = 1$ head.
- For TH, $x = 1$ head.
- For TT, $x = 0$ heads.

The four outcomes of tossing the coins are equally likely, each with a probability $\frac{1}{4}$. We can make a table showing the possible outcomes for the number of heads observed $x = 0$, 1, and 2 along with their associated probabilities.



Now multiply each value of x by its probability and add the products:

$$\begin{aligned}\mu &= \sum_{i=1}^{n} x_i \cdot P(x_i) \\ &= \left(0 \cdot \frac{1}{4}\right) + \left(1 \cdot \frac{2}{4}\right) + \left(2 \cdot \frac{1}{4}\right) \\ &= 0 + \frac{2}{4} + \frac{2}{4} \\ &= 1\end{aligned}$$

If this experiment is performed many, many times, we expect an average of 1 head per toss.

Example 6.5.6

A dart is equally likely to land anywhere on the dartboard. A player earns 40 points for landing on gray, 20 points for landing on yellow, and 0 points for landing on black.



On average, how many points will a player earn per throw?

Solution

Notice that the chance of landing on black is different than the chances of landing on yellow and gray. Of the 16 squares, 5 are gray, 5 are yellow, and 6 are black. We can make a table showing the number of points earned and probabilities for each outcome.



Now multiply each value of x by its probability and add the products:

$$\begin{aligned}\mu &= \sum_{i=1}^{n} x_i \cdot P(x_i) \\ &= \left(40 \cdot \frac{5}{16}\right) + \left(20 \cdot \frac{5}{16}\right) + \left(0 \cdot \frac{6}{16}\right) \\ &= \frac{200}{16} + \frac{100}{16} + 0 \\ &= \frac{300}{16} \\ &= 18.75\end{aligned}$$

A player earns an average of 18.75 points per throw over the long run. It is important to understand that a player never earns exactly 18.75 points. If the player throws the dart many times, keeps track of the scores, and averages all the scores together, it should be close to 18.75 points.

You Try It $6.5.C$

In a carnival game, a player tosses a bean sack toward three concentric circles on a target board. The bean sack is not equally likely to land in each circle. Based on past records, the probabilities are:

- Landing in the blue circle: 50%
- Landing in the red circle: 30%
- Landing in the green circle: 20%

A player earns 15 points for landing in the blue circle, 5 points for landing in the red circle, and 0 points for landing in the green circle.

Find the expected number of points a player earns per toss.

Answer available at the end of this section.

Positive, Negative, and Zero Expected Value

Expected value plays a crucial role in gaming and gambling. In general, if the expected value of your winnings in a money-based game is negative, the game is unfavorable -- you should expect to lose money on average. A game with a positive expected value would be advantageous, though such games are extremely rare outside of carefully designed strategies.

It is important to remember that even when a game has a positive expected value, most players might still lose money while a single fortunate player wins a large amount. Expected value describes the long-run average, not the outcome of any individual play.

When the expected value of a game is 0, the game is considered **fair**, meaning neither player has a mathematical advantage.

Example 6.5.7

A carnival game consists of drawing one ball from a box containing two yellow balls, five red balls, and eight white balls. If the ball is yellow, you win $5. If the ball is red, you win $2. If the ball is white, you lose $3.

Is this a fair game? If you play the game many times, how much would you expect to gain or lose per game?

Solution

It is helpful to make a table that summarizes all the information before finding the expected value.



Now use the formula for the expected value:

$$\mu = \$5\left(\frac{2}{15}\right) + \$2\left(\frac{5}{15}\right) + -\$3\left(\frac{8}{15}\right) \approx -\$0.27$$

Since the expected value is not zero this is not a fair game. The expected value is negative which indicates that a player can expect to lose about 27 cents per game on average when playing this game many times. That means that the carnival will make an average of $0.27 for every game played.

Picking a ball from the bag is a random experiment so we cannot predict what will happen if we play the game once. We can predict what will happen only if we play the game many times.

You Try It 6.5.D

A game consists of rolling a colored die with three green sides, two red sides, and one blue side. If you roll green, you earn 2 points. A roll of a red earns you 1 point. If you roll blue, you lose 8 points.

Is this a fair game? If not, how many points can you expect to earn or lose per game?

Answer available at the end of this section.

Example 6.5.8

You purchase a raffle ticket for a charity fundraiser. Each ticket costs $5, and the charity sells 2,000 tickets. One ticket will be drawn, and the winner receives a prize worth $4,000. Compute the expected value of your gain or loss when participating in this raffle.

Solution

We want the long-run average amount a player gains or loses. If you buy one ticket, there are two possible outcomes:

- You win the raffle with probability $\frac{1}{2000}$.
- You lose with probability $\frac{1999}{2000}$.

Remember that you pay $5 whether you win or lose.

- If you win, your net gain is $\$4000 - \$5 = \$3995$.
- If you lose, your net gain is $-\$5$.



Now multiply each outcome by its probability and add the products:

$$\begin{aligned}\mu &= \sum_{i=1}^{n} x_i \cdot P(x_i) \\ &= \left(3995 \cdot \frac{1}{2000}\right) + \left(-5 \cdot \frac{1999}{2000}\right) \\ &= 1.9975 + -4.9975 \\ &= -3\end{aligned}$$

The expected value of a ticket in this raffle is –$3.00. On average, participants lose $3 by entering. No one will ever lose exactly $3, but this represents the long-term average loss across all players. Most people will lose their $5 entry fee, while one lucky person will win the prize. In this case, the loss may be acceptable since the money supports a charity.

Example 6.5.9

A friend offers you a chance to play a game in which you pay $2 to roll two standard dice. If the dice show different values, you lose your money. If the dice match (that is, you roll doubles), you receive $5 back. Is this a fair game?

Solution

To determine whether the game is fair, we compute the expected value of your winnings.

There are $6 \times 6 = 36$ possible outcomes when rolling two dice.

- Six of these outcomes are doubles: (1,1), (2,2), ..., (6,6). In these cases, your net winnings are $\$5 - \$2 = \$3$. (You win $5 but you paid $2 to play.)
- The remaining 30 outcomes are not doubles. In these cases, you lose your $2, or $-\$2$.



Now multiply each possible gain or loss by its probability and add the results:

Now use the formula for the expected value:

$$\mu = (\$3 \cdot \tfrac{6}{36}) + (-\$2 \cdot \tfrac{30}{36}) = -\$0.50$$

Your expected value for this game is –$0.50. If you play many, many times, your average loss will be 50 cents per game. You will sometimes win $5 and sometimes lose $2, but over the long run the average settles to a loss of $0.50.

Not surprisingly, the expected value for casino games is always negative for the player and therefore positive for the casino. The casino's expected value must be positive, or it would go out of business. Players just need to keep in mind that when they play a game repeatedly, their expected value is negative. That is fine so long as you enjoy playing the game and think it is worth the cost. But it would be wrong to expect to come out ahead. If you are ever ahead, you might want to consider walking away from the game.

Likewise, an insurance company makes sure its expected return on policies sold will be positive. That means, you can expect to have a negative return when you buy a policy.

Example 6.5.10

The cost of a $50,000 life insurance policy is $150 per year for a person who is 21 years old. Assume the probability that a person will die at age 21 is 0.001.

What is the insurance company's expected return on selling this type of policy?

Solution

There are two possible outcomes:

- The person lives during the year (probability 0.999).
- The person dies during the year (probability 0.001).

From the company's perspective:

- If the person lives, the company keeps the $150 premium as profit.
- If the person dies, the company must pay the $50,000 benefit. Its net gain is $\$150 - \$50,000 = -\$49,850$.



The expected value for one policy is

$$\mu = (150 \cdot 0.999) + (-\$49,850 \cdot 0.001) = \$100.$$

When the company sells a very large number of these policies, it will average a profit of $100 on each one. Most customers will live, generating a profit of $100, while a small number will die, resulting in a large payout.

You Try It 6.5.E

You take out a fire insurance policy on your home. The yearly premium is $250. In case of fire, the insurance company will pay you $200,000. The probability of a house fire in your area is 0.003%. What is the insurance company's expected value for selling one of these policy?

Answer available at the end of this section.

Section Summary

Expected value describes the long-term average numerical outcome of a chance process. This expected value (or mean) is denoted by μ and is computed using the formula

$$\mu = \sum_{i=1}^{n} x_i \cdot P(x_i).$$

This value represents what you would expect on average if the experiment were repeated many, many times. You may never observe the expected value in a single trial, but it reflects the mean of all possible outcomes weighted by their probabilities.

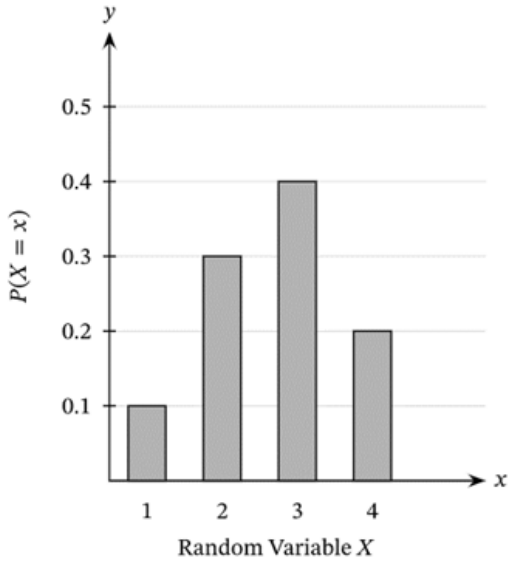
Expected value is a powerful tool for interpreting and evaluating real-world situations involving chance. It allows us to determine whether a game is fair, whether one player has an advantage, and whether a business model -- such as an insurance policy or a gambling operation -- is profitable over the long run. In every case, the expected value provides a clear numerical measure of the long-term behavior of a random process.

Section 6.5 Review Questions

1. Consider the outcomes x and their corresponding probabilities shown in the table. Find the expected value of x.



2. Consider the outcomes x and their corresponding probabilities shown in the bar graph. Find the expected value of x.



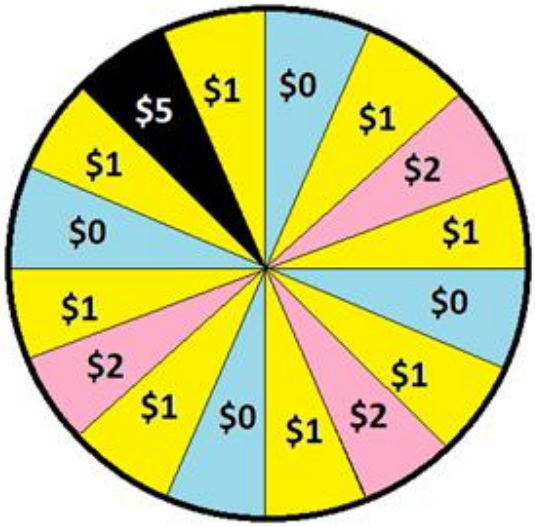
3. A friend offers you the opportunity to win money by playing her card game. Find the expected amount you win or lose per game. The rules are
    - You must pay her $2 to play.
    - If you pick a spade from a shuffled pack, you win $9.
    - If you pick any other card, you lose your money.

4. An insurance company will sell a $10,000 one-year term life insurance policy to an individual in a particular risk group for a premium of $368. Find the expected profit to the company of a single policy if a person in this risk group has a 98% chance of surviving one year.

5. For a certain tutor the probabilities for various numbers of student no-shows per day are shown in the following table. What is the expected value for the number of no-shows per day?



6. At the carnival, you pay $2 to spin the wheel below. The vendor pays you the amount you spin. Assume the sectors on the wheel are equally likely.



   a. What is the expected value of *your gain or loss* when playing the game?
   b. What is the expected value of the *amount you spin* when playing the game?
   c. Why are the answers to parts (a) and (b) different?

7. A company estimates that 4% of their products will fail within 2 years of the purchase. It costs $500 to replace the product. The company decides to sell a 2-year service plan for $48 that will replace the product for free. On average, how much can the company expect to earn or lose per service plan sold?

8. Marina is playing a game in which she needs to throw a ball into a bucket. If she throws the ball into the bucket, she wins $100. If not, she wins nothing. The odds against Maria throwing the ball into the bucket is 3 to 2. What is the expected value of Marina's winnings?

9. An insurance company insures a house worth $250,000 for an annual premium of $500. If the probability of the house being destroyed is .0015 and assuming either total loss or no loss, what is the insurance company's expected annual profit or loss for the policy?

10. A game consists of rolling a colored die with three green sides, two red sides, and one blue side. A roll of green pays you $2. A roll of a red loses your money. A roll of blue pays you $6. What is a "fair" price to pay to play?

Answers

You Try It $6.5.A$
1.35 toppings

You Try It $6.5.B$
20,550 people

You Try It $6.5.C$
9 points

You Try It $6.5.D$
This is a fair game because the expected value is 0 points.

You Try It $6.5.E$
$244

This page titled 6.5: Expectation Involving Chance is shared under a license and was authored, remixed, and/or curated by Andy Jones.

- 12.6: Expected Value by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.
- 3.4: Expected Value and Law of Large Numbers 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.

7: COUNTING TECHNIQUES

7.1: The Fundamental Counting Principle
7.2: Permutations and Combinations
7.3: Probability Using Counting Techniques

This page titled 7: Counting Techniques is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.1: The Fundamental Counting Principle

Learning Objectives

- Apply the *Fundamental Counting Principle* to count outcomes.
- Calculate ordered arrangements using factorials.

Introduction

Counting may seem as simple as 1-2-3 -- you already know how to count, or you wouldn't be taking a college-level math class. But in this section, our focus is not on *basic* counting. Instead, we study efficient methods for counting large numbers of outcomes, especially when listing them one by one would be impossible.

A key part of this work is learning what to count and when two outcomes should be considered different. Once we can identify distinct outcomes and count them accurately, we can apply these techniques directly to probability. In fact, the ultimate goal of mastering these counting methods is to integrate them into the study of probability and to analyze chance processes more effectively.

Fundamental Counting Principle

The simplest of the counting methods is the ***Fundamental Counting Principle*** **(FCP).** This method is generally used when there are choices that must be made in succession and there are several options for each choice.

Example 7.1.1

Gretchen is remodeling her kitchen. For her design package, she must choose a floor, a counter, and a sink from the options shown below. She wants to list all possible kitchen designs she can create.

Gretchen's Kitchen Design Options

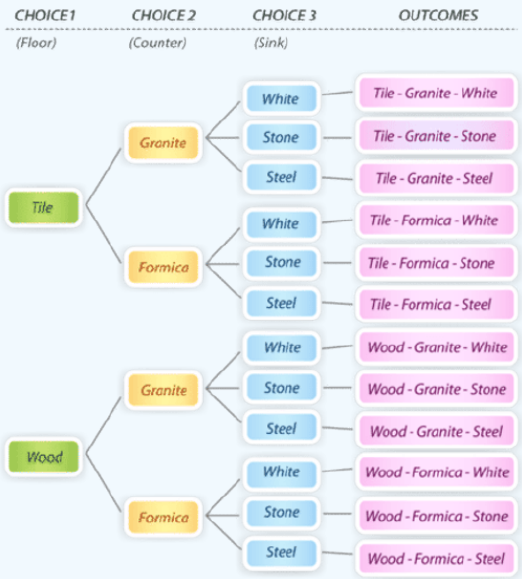
- Choice 1: Floor — Tile, Wood
- Choice 2: Counter — Granite, Formica
- Choice 3: Sink — White, Steel, Stone

A tree diagram is a helpful tool for visualizing all possible combinations. It organizes outcomes clearly and ensures that none are missed.

Begin the tree diagram by branching from each floor option (tile or wood) to each counter option (granite or formica.) Every counter choice must appear for each floor choice, producing $2 \times 2 = 4$ floor-counter options.



Next, extend the diagram by branching each of these four floor-counter options to the three sink options. Each sink choice must repeat for every floor-counter pair.



The completed tree diagram displays $4 \times 3 = 12$ distinct kitchen designs, each representing one unique combination of floor, counter, and sink.

Tree diagrams are useful for visualizing choice options and outcomes, but they can become time-consuming when there are many choices or many stages in the decision process. A faster method for determining the total number of possible outcomes is to multiply the number of options available at each stage. Using multiplication, the total number of possible design packages is $2 \times 2 \times 3 = 12$.

Fundamental Counting Principle

If there are n possible outcomes for event A and m possible outcomes for event B, then there are $n \cdot m$ total possible outcomes for event A followed by event B.

This principle can be generalized for three, four, or even more events.

Example 7.1.2

A restaurant offers a special dinner menu every day. A diner must choose from three entrees, five appetizers, and four desserts. A customer can only select one item from each category. How many different meals can be ordered from the special dinner menu?

Solution

To find the total number of different meals a customer can order, apply the *Fundamental Counting Principle* and multiply the number of choices in each category:

$$\underline{3} \times \underline{5} \times \underline{4} = 60$$

There are 60 different meals that can be ordered from the special dinner menu.

You Try It 7.1.A

When purchasing a computer, the e-Box laptop computer offers customers several different options for screen, memory, and color as shown below. How many ways can a customer choose to personalize her selection of a new computer?

- **Screen:** small, medium, or large
- **Memory:** standard, 1 GB, or 2 GB
- **Color:** pearl, gray, blue, or black

Answer available at the end of this section.

The *Fundamental Counting Principle* may seem like a very simple idea, but it is very powerful. Many complex counting problems can be solved using this strategy.

Example 7.1.4

A quiz contains 4 true-or-false questions. In how many different ways can a student answer the quiz?

Solution

Each question has 2 possible answers (True or False). Treat each question as a separate event:

- Question 1: 2 options
- Question 2: 2 options
- Question 3: 2 options
- Question 4: 2 options

Using the *Fundamental Counting Principle*, multiply the number of options for each question:

$$\underline{2} \cdot \underline{2} \cdot \underline{2} \cdot \underline{2} = 2^4 = 16.$$

There are 16 different ways a student can answer the four true-or-false questions.

You Try It 7.1.B

A student flips a coin 5 times. Each flip can result in Heads or Tails. In how many different ways can the sequence of 5 flips occur?

Answer available at the end of this section.

Example 7.1.5

How many 5-digit ZIP codes can be formed if the ZIP code cannot begin with 0?

Solution

A ZIP code has five digits, so we consider five separate choice events.

- The first digit cannot be 0. It has 9 possible options (1 through 9).
- Each of the remaining four digits can be any digit from 0 through 9, giving 10 options for each position.

Using the *Fundamental Counting Principle,* multiply the number of choices for each digit:

$$\underline{9} \times \underline{10} \times \underline{10} \times \underline{10} \times \underline{10} = 9 \times 10^4 = 90,000$$

There are 90,000 possible 5-digit ZIP codes that do not begin with 0.

You Try It 7.1.C

A company manufactures electronic devices, each labeled with a 5-character serial number. The first character must be a letter from A–F, and the remaining four characters are digits from 0–9. How many different serial numbers can the company create?

Answer available at the end of this section.

Example 7.1.6

How many different 4-digit student ID numbers can be made if the first digit must be a 5 or 6 and repetitions are not permitted?

Solution

A student ID number has four digits, so we treat each digit as a separate choice event.

- The first digit has 2 possible options (5 or 6).
- The second digit can be any digit except the first, giving 9 options.
- The third digit must differ from both the first and second digits, leaving 8 options.
- The fourth digit must differ from the first three digits, leaving 7 options.

Using the *Fundamental Counting Principle*, multiply the number of choices for each position:

$$\underline{2} \times \underline{9} \times \underline{8} \times \underline{7} = 1,008$$

There are 1,008 different possible student ID numbers under these conditions.

Example 7.1.7

Some license plates in Maryland consist of three letters (A to Z) followed by three digits (0 to 9). Count how many license plates of this type are possible when

a. letters and digits may be repeated.

Solution

$$\underbrace{(26 \cdot 26 \cdot 26)}_{\text{letters}} \cdot \underbrace{(10 \cdot 10 \cdot 10)}_{\text{digits}} = 17,576,000 \text{ license plates}$$

b. letters can be repeated but digits cannot.

Solution

$$\underbrace{(26 \cdot 26 \cdot 26)}_{\text{letters}} \cdot \underbrace{(\underline{10} \cdot \underline{9} \cdot \underline{8})}_{\text{digits}} = 12,654,720 \text{ license plates}$$

c. both digits and letters can be repeated but the first digit cannot be zero.

Solution

$$\underbrace{(26 \cdot 26 \cdot 26)}_{\text{letters}} \cdot \underbrace{(\underline{9} \cdot \underline{10} \cdot \underline{10})}_{\text{digits}} = 15,818,400 \text{ license plates}$$

d. all letters and digits can be used but cannot be repeated.

Solution

$$\underbrace{(\underline{26} \cdot \underline{25} \cdot \underline{24})}_{\text{digits}} \cdot \underbrace{(\underline{10} \cdot \underline{9} \cdot \underline{8})}_{\text{letters}} = 11,232,000 \text{ license plates}$$

You Try It 7.1.D

How many 3-digit area codes can be formed where the first and last digits are odd, and digits cannot be repeated?

Answer available at the end of this section.

Factorial

Consider the following situation:

Four customers arrive at a grocery store checkout at the same time. The cashier can serve only one person at a time. In how many different ways can the four people line up to pay for their items?

This is a classic ordering problem. Each arrangement of the four customers represents a different possible line order. To build the line, think of the process step by step:

1. First position: Any of the 4 customers could be first.
2. Second position: After the first person is chosen, 3 customers remain.
3. Third position: After two people are placed, 2 customers remain.
4. Fourth position: Only 1 customer is left for the final position.

Using the *Fundamental Counting Principle,* multiply the number of choices at each stage: $4 \times 3 \times 2 \times 1 = 24$. So there are 24 different ways the four customers can line up.

A product similar to $4 \times 3 \times 2 \times 1$ appears so often in counting problems that mathematicians use special notation and a name for it -- a **factorial.**

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

The symbol $4!$ is read as "*four factorial*."

Factorial

Suppose n represents a counting number. Then, $n!$ and $0!$ are defined as follows:

$$n! = n \cdot (n-1) \cdot (n-2) \ldots 3 \cdot 2 \cdot 1$$

$$0! = 1$$

Factorials can be generalized to any counting number. For example, if there were 7 customers in line, then there are $7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5,040$ ways to line them up in order. The factorial function is especially important in counting problems where order matters, such as arranging books on a shelf, assigning seats, or ordering digits in a code. Any time you want to know how many different ways you can place n distinct items in order, the answer is $n!$.

You may be wondering about special case $0! = 1$ in the definition of factorial above. This may seem surprising at first. However, this value is defined so that counting formulas work consistently even when no objects are being arranged. And it makes sense to say that if you had 0 customers waiting in line, there is only one way you could arrange the 0 customers. That is, there is one way to do nothing.

Example 7.1.8

Evaluate $5!$, $9!$, and $12!$.

Solution

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

$9! = 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 40,320$

$12! = 12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 479,001,600$

Example 7.1.9

Five musicians arrive backstage for a concert, and the stage manager needs to decide the order in which they will perform their solo pieces. Each musician must take a unique position in the lineup, and each muscian plays exactly once.

In how many different ways can the five musicians be arranged for the performance?

Solution

The number of possible ordered arrangements of 5 distinct objects is $5!$:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

So, there are 120 different performance orders the stage manager could choose.

You Try It 7.1.E

Six athletes qualify for the final round of a track-and-field event. The officials need to determine the order in which the athletes will compete.

In how many different ways can the six athletes be arranged for the competition order?

Answer available at the end of this section.

Factorials grow get very large, very fast. It can be unpleasant to type so many factors when n is large. Most calculators have a factorial command. Here is how to calculate a factorial on the TI calculator.

Factorials on the TI-83/84 Calculator

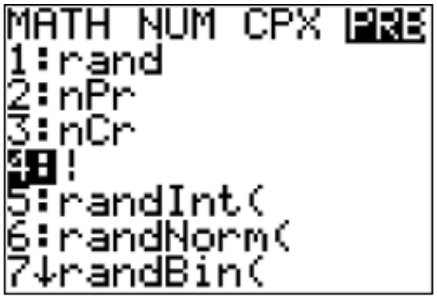
To find $n!$, type the value of n first.

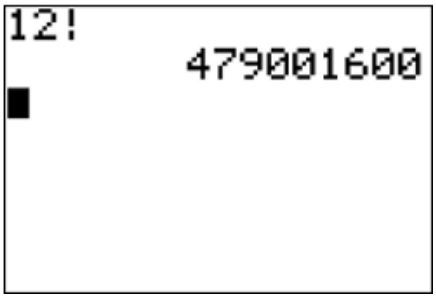
Press [MATH] and use the arrow to scroll to the right to the PRB (probability) menu.

Select option 4: ! and then press ENTER.

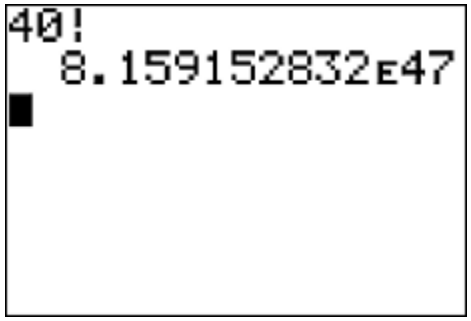
The screenshots below show the menu and the proper procedure for finding 12!.







A calculator will express factors with large values in scientific notation (often rounded.) As shown below, the TI calculator finds $40!$ to be approximately 8.16×10^{47}:



So, $40!$ is about 816 followed by 45 zeros. This is only an approximate value because the calculator has truncated the value to 9 places to the right of the decimal point in scientific notation.

Permutations

Example 7.1.9 about arranging musicians highlights a common type of counting problem called a permutation. **A permutation**

is an ordered arrangement of objects where the objects are not repeated. The essential feature of a permutation is that order matters. Changing the order creates a different outcome.

To see this idea in action, consider the three letters C, A, and T. Because order matters, each distinct sequence of these letters counts as a different permutation. There are exactly six such ordered arrangements:

ACT ATC CAT CTA TAC TCA.

If we add a fourth letter, say S, the number of possible ordered arrangements increases dramatically. There are now 24 different four-letter permutations of C, A, T, and S:

ACST	CAST	SACT	TACS
ACTS	CATS	SATC	TASC
ASCT	CSAT	SCAT	TCAS
ASTC	CSTA	SCTA	TCSA
ATCS	CTAS	STAC	TSAC
ATSC	CTSA	STCA	TSCA

Counting these ordered arrangements is the same type of problem as counting the number of ways four customers can line up at a checkout stand. In both situations, we are arranging four different objects in order. The number of possible arrangements is

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24.$$

Counting Permutations of n Items

In general, the number of ways to order n different objects is given by the factorial expression $n!$.

Factorials provide a powerful and compact way to count ordered arrangements, and they form the foundation for many topics in combinatorics and probability. We will use permutations in the next section.

Example 7.1.10

How many seating arrangements are there for 8 guests who attend a dinner party?

Solution

We want to count the number of ways that 8 different objects (guests) that can be assigned to 8 positions (seats). Changing where a guest sits creates a different seatig arrangement so this is a permutation. The number of permutations using all 8 objects is $8! = 40,320$ seating arrangements.

It should be noted that Example 7.1.10 also can be solved using the *Fundamental Counting Principle*. For the first seat, there are 8 choices of people. For the second seat, there are 7 remaining choices, since one person has already been seated. For the third seat, there are 6 choices, since two people are already seated. By the time we get to the last seat, there is only one person left. Therefore, using the *Fundamental Counting Principle*, we get 8 × 7 × 6 × 5 × 4 × 3 × 2 × 1 = 40,320 arrangements. In fact, any permutation can be solved using the *Fundamental Counting Principle*.

Section Summary

The *Fundamental Counting Principle* states that if an event A can occur in n ways and a following event B can occur in m ways, then the sequence "event A followed by event B" can occur in $n \cdot m$ different ways. This principle extends to any number of

consecutive events by multiplying the number of outcomes for each event.

When n is a counting number, we define the factorial of n as the product of all positive integers from n down to 1:

$$n! = n \cdot (n-1) \cdot (n-2) \ ... \ 3 \cdot 2 \cdot 1.$$

By definition, $0! = 1$.

A permutation is an ordered arrangement of distinct items, with no repetitions allowed. Because order matters, each unique ordering counts as a different permutation. If there are n distinct objects to be arranged, then the number of possible ordered arrangements is $n!$.

Section 7.1 Review Questions

1. Tilly spins a spinner that has a red, a blue, and a yellow section and then tosses a single six-sided die. How many different outcomes are possible for spinning and then tossing?

2. Flying to Los Angeles from Washington, DC, you can choose one of three airlines and can also choose either first class or coach. Based on these choices, how many travel options do you have to LA?

3. When purchasing a computer, the e-Box laptop computer offers customers several different options for screen, memory, and color as shown below. How many ways can a customer choose to personalize her selection of a new computer?

    - **Screen:** small, medium, or large;
    - **Memory:** standard, 1 GB, 2 GB, 4GB, or super 8GB
    - **Color:** pearl, gray, blue, black.

4. You are taking an 8-question survey on your experience at Taco Bell. For the first 5 questions you can answer *Below Average, Average*, or *Above Average* for each question. The last three questions you can respond with either *Agree* or *Disagree*. How many different ways could you complete this survey?

5. In how many ways can you roll a regular six-sided die twice and get an even number first followed by a number less than 5?

6. An automobile license plate is designed to show a letter of the English alphabet, followed by a five-digit number. Find the number of different license plates that can be issued if ...
    a. digits can be repeated.
    b. digits cannot be repeated.

7. A social security number is made of 9 digits. The first digit cannot be 0. How many different social security numbers can be created if...
    a. digits can be repeated?
    b. digits cannot be repeated?

8. How many ways can 7 video game cartridges be arranged on a shelf from left to right?

9. The school orchestra is planning to play 8 pieces of music at their next concert. In how many different ways can the pieces of music be sequenced in the program?

10. How many 3-digit area codes can be formed if the first and last digits are odd, and digits cannot be repeated?

11. Determine the number of sequences of heads and tails are possible when a coin is tossed
    a. twice
    b. three times
    c. five times
    d. n times (*Hint:* Look for a pattern in the results of a through c.)

Answers

You Try It $7.1.A$
36 ways

You Try It $7.1.B$
$2^5 = 32$ ways

You Try It $7.1.C$
60,000 serial numbers

You Try It $7.1.D$
180 area codes

You Try It $7.1.E$
$6! = 720$ ways

This page titled 7.1: The Fundamental Counting Principle is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- 3.5: Counting Methods 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.

7.2: Permutations and Combinations

Learning Objectives

- Distinguish between permutations and combinations in applications.
- Count permutations and combinations using formulas.
- Use technology to evaluate counting expressions.
- Solve problems involving multiple counting methods.
- Count distinguishable permutations when some items are identical.

Introduction

Counting problems often involve determining how many different outcomes are possible without listing them all. Two of the most important tools for doing this are permutations and combinations.

Permutations deal with situations where order matters such as in arranging books on a shelf, assigning seats, or ranking contestants. Combinations apply when order does not matter such as in forming committees, choosing teams, or selecting items from a set. Both ideas build on the *Fundamental Counting Principle* and the use of factorial notation, allowing us to count large sets of possibilities efficiently and systematically.

More on Permutations

At the end of Section 7.1, we counted permutations -- ordered arrangements of n distinct items. When all n items are used, the number of possible arrangements is $n!$. For example, if we have 8 pieces of artwork to hang, then there are $8! = 40,320$ different ways to arrange all eight pieces in a row.

Now consider a different situation: Suppose the wall has only 5 hooks. We still have 8 pieces of artwork, but we can choose and display only 5 of them. How many different ordered arrangements of 5 pieces can be made?

Counting Permutations Directly with the Fundamental Counting Principle

This seems like a familiar type of problem. We can use the *Fundamental Counting Principle*. Because there are only 5 hooks, we fill them one at a time:

- **1st hook:** 8 choices
- **2nd hook:** 7 choices remain
- **3rd hook:** 6 choices remain
- **4th hook:** 5 choices remain
- **5th hook:** 4 choices remain

Multiplying these outcomes gives $8 \times 7 \times 6 \times 5 \times 4 = 6,720$ ways to choose and hang 5 pieces of art when there are 8 pieces to choose from.

Connecting to Permutations

This is still a permutation problem because we are counting ordered arrangements of distinct objects with no repetition. The difference is that we are not using all 8 pieces -- only 5 of them.

We cannot use $8!$, because that would count arrangements of all 8 pieces. Instead, observe that

$$8 \times 7 \times 6 \times 5 \times 4 = \frac{8 \times 7 \times 6 \times 5 \times 4 \times (3 \times 2 \times 1)}{(3 \times 2 \times 1)} = \frac{8!}{(8-5)!}.$$

This leads to the general formula for permutations of n objects chosen r at a time:

$$ {}_nP_r = \frac{n!}{(n-r)!}. $$

While the *Fundamental Counting Principle* always works, the permutation formula is often more efficient -- especially when the numbers become large or when many positions must be filled.

Permutation

A **permutation** is an arrangement of distinct items in which the order of the items matters.

The number of permutations that can be formed from n objects when all n of them are used is given by the formula

$$ {}_nP_n = n! $$

The number of permutations that can be formed from n objects when only r of them are taken at a time is given by the formula

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

Example 7.2.1

There are 10 cars in a race. Determine how many ways three cars can be awarded 1st, 2nd, and 3rd place.

Solution

Because we are counting finishing orders, this is a permutation problem. We are considering a total of $n = 10$ objects, and we wish to choose and place 3 objects into ordered positions $r = 3$.

Using the permutation formula,

$$ {}_nP_r = \frac{n!}{(n-r)!} \qquad \text{Write the permutation formula.} $$

$$ {}_{10}P_3 = \frac{10!}{(10-3)!} \qquad \text{Substitute } n = 10 \text{ and } r = 3. $$

$$ = \frac{10!}{7!} \qquad \text{Simplify the denominator.} $$

$$ = \frac{10 \times 9 \times 8 \times \cancel{7!}}{\cancel{7!}} \qquad \text{Expand numerator to isolate 7!,} $$

$$ = 10 \times 9 \times 8 \qquad \text{These factors remain after cancellation.} $$

$$ = 720 \qquad \text{Multiply to find final product.} $$

There are 720 different ways to award 1st, 2nd, and 3rd place among the 10 cars.

The next-to-last line matches exactly what we obtain using the *Fundamental Counting Principle:* $10 \times 9 \times 8 = 720$.

Example 7.2.2

The board of directors at a foundation has 13 members. Three officers will be elected to the positions of president, general director, and treasurer.

How many ways could these positions be filled?

Solution

There are three positions to fill, and each position is different. This means we are counting permutations rather than simply choosing a group of three people. For example, Mr. Smith, Mr. Hale, and Ms. Osborn could fill the positions in many different orders, each representing a different outcome. Thus, the assignment of individuals to specific offices is an ordered arrangement.

There are $n = 13$ board members and $r = 3$ officer positions.

Using the permutation formula,

$$
\begin{aligned}
{}_{13}P_3 &= \frac{13!}{(13-3)!} && \text{Substitute } n = 13 \text{ and } r = 3.\\
&= \frac{13!}{10!} && \text{Simplify the denominator}\\
&= \frac{13 \times 12 \times 11 \times \cancel{10!}}{\cancel{10!}} && \text{Expand numerator to isolate } 10!.\\
&= 13 \times 12 \times 11 && \text{Cancel like factors.}\\
&= 1,716 && \text{Multiply to find final product.}
\end{aligned}
$$

Therefore, there are 1,716 different ways to elect the three officers.

This matches what we would obtain using the *Fundamental Counting Principle*: 13 choices for president, 12 remaining choices for general director, and 11 remaining choices for treasurer gives $13 \times \ 12 \times 11 = 1,716$.

Example 7.2.3

A music festival has 9 performers available. For the evening concert, the organizer must choose 4 performers and schedule them in a specific performance order. No performer may appear more than once in the lineup.

How many different ordered lineups are possible?

Solution

Because the performance order matters and no performer is repeated, this is a permutation problem. We have $n = 9$ performers and $r = 4$ positions in the lineup.

$$
\begin{aligned}
{}_9P_4 &= \frac{9!}{(9-4)!} && \text{Substitute } n = 9 \text{ and } r = 4 \text{ into } \frac{n!}{(n-r)!}.\\
&= \frac{9!}{5!} && \text{Simplify the denominator}\\
&= \frac{9 \times 8 \times 7 \times 6 \times \cancel{5!}}{\cancel{5!}} && \text{Expand numerator to isolate } 5!.\\
&= 9 \times 8 \times 7 \times 6 && \text{Cancel like factors.}\\
&= 3,024 && \text{Multiply to find final product.}
\end{aligned}
$$

There are 3,024 possible ordered concert lineups.

You Try It 7.2.A

Use the permutation formula to find ${}_{15}P_5$.

Answer available at the end of this section.

You Try It 7.2.B

A culinary institute is hosting a cooking showcase with 10 professional chefs available. The event features 4 demonstration stations, each requiring a different cooking specialty.

In how many ways can the 4 stations be assigned to the chefs?

Answer available at the end of this section.

Combinations

We now turn to a closely related situation: selecting r items from n items when the order of selection does not matter. In this setting, rearranging the same chosen items does not produce a new group. A selection of items without repetition and without regard to order is called a **combination**.

Developing a Formula for Combinations from the Permutation Formula

To gain insight into a formula for counting combinations, let's explore a particular problem.

Suppose we have five students -- Andy (A), Barry (B), Cheryl (C), Darren (D), and Ellen (E). We want to select three students to form a team. Each distinct three-student team is a combination, because the order in which the students are listed does not change the team. Our goal is to determine how many different three-student teams can be formed.

However, ***if order did matter***, then selecting three students would be a permutation:

$$ {}_5P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = 60. $$

This value counts every possible ordered arrangement of three students. But, when f*orming a team*, the order of the students is irrelevant. Therefore, we have overcounted the number of teams.

To correct this, determine how many times each unordered team has been counted. Consider the team consisting of students A, B, and C. If order mattered, the following six permutations would all be counted separately:

ABC ACB BAC BCA CAB CBA.

Yet all six of these permutations represent the same team. These six arrangements correspond to the number of ways to order three distinct objects: $3! = 3 \times 2 \times 1 = 6$.

Each unordered team appears six times in the permutation count. To obtain the number of combinations, we divide 60 permutations by the number of arrangements that form the same team $(3! = 6)$.

$$ {}_5C_3 = \frac{{}_5P_3}{3!} = \frac{60}{6} = 10. $$

So, one way to count combinations is to first count permutations ${}_5P_3$ and then divide by $3!$.

We can also express this calculation more directly using factorials:

$$
\begin{aligned}
{}_5C_3 &= \frac{{}_5P_3}{3!} \\
&= \frac{\left(\frac{5!}{(5-3)!}\right)}{3!} \\
&= \frac{5!}{(5-3)! \times 3!} \\
&= \frac{5!}{2! \times 3!} \\
&= \frac{5 \times 4 \times \cancel{3!}}{(2 \times 1) \times \cancel{3!}} \\
&= \frac{20}{2} \\
&= 10
\end{aligned}
$$

General Formula for Combinations

This reasoning leads to the general formula for counting combinations:

$$
{}_nC_r = \frac{n!}{(n-r)!\ r!}
$$

The extra factor of $r!$ in the denominator of the combination formula accounts for the number of ways to order the selected r items -- order that combinations intentionally ignore. This formula counts the number of ways to choose r items from n items without repetition and without regard to order.

Combination

A **combination** is a selection of distinct items in which the arrangement or order of those items is ignored.

The number of combinations that can be formed from n objects when only r of them are taken at a time is given by the formula

$$
{}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!\ r!}
$$

Example 7.2.4

How many different groups of 3 club members can be selected as delegates to a convention if the club has 13 members?

Solution

Because the delegates form a group and the order of selection does not matter, this is a combination problem involving 13 members taken 3 at a time.

Using the combination formula ${}_{12}C_3$:

$\frac{12!}{(12-3)!\ 3!}$	Substitute into the combination formula.
$= \frac{12!}{9!\ 3!}$	Simplify the denominator subtraction.
$= \frac{12 \times 11 \times 10 \times \cancel{9!}}{\cancel{9!}\ 3!}$	Expand numerator to isolate 9!.
$= \frac{12 \times 11 \times 10}{3!}$	Cancel the matching factorials.
$= \frac{12 \times 11 \times 10}{3 \times 2 \times 1}$	Expand the remaining factorial 3!.
$= \frac{1320}{6}$	Simplify the numerator and denominator.
$= 220$	Divide to find the final product.

There are 220 ways to select 3 delegates for the convention.

Example 7.2.5

You are taking an education course that requires you to read 4 books from a list of 10. You are free to select any 4 books and read them in whichever order that pleases you.

How many different sets of 4 books can you choose?

Solution

Because you are forming a group of books and the order of selection does not matter, this is a combination problem.

Use the combination formula for ${}_{10}C_4$:

$\frac{10!}{(10-4)!\ 4!}$	Substitute into the combination formula.
$= \frac{10!}{6!\ 4!}$	Simplify the denominator subtraction.
$= \frac{10 \times 9 \times 8 \times 7 \times \cancel{6!}}{\cancel{6!}\ 4!}$	Expand numerator to isolate 6!.
$= \frac{10 \times 9 \times 8 \times 7}{4!}$	Cancel the matching factorials.
$= \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1}$	Expand the remaining factorial 4!.
$= 210$	Simplify and find the final product.

There are 210 different ways to select 4 books from the reading list.

You Try It 7.2.C

Use the combination formula to find ${}_{11}C_3$.

Answer available at the end of this section.

You Try It $7.2.D$

A dessert shop is creating a weekend ice cream cone special featuring 3 different flavors chosen from its menu of 14 flavors.

How many different selections could a customer make?

Answer available at the end of this section.

Using Technology for Permutations and Combinations

As we move into the topic of using technology to count permutations and combinations, it is crucial to remember that technology can only produce correct results when *we* choose the correct counting method. Modern calculators can evaluate formulas instantly, but they cannot determine whether a situation requires order-sensitive arrangements or order-insensitive selections. That distinction must come from your understanding of the scenario itself.

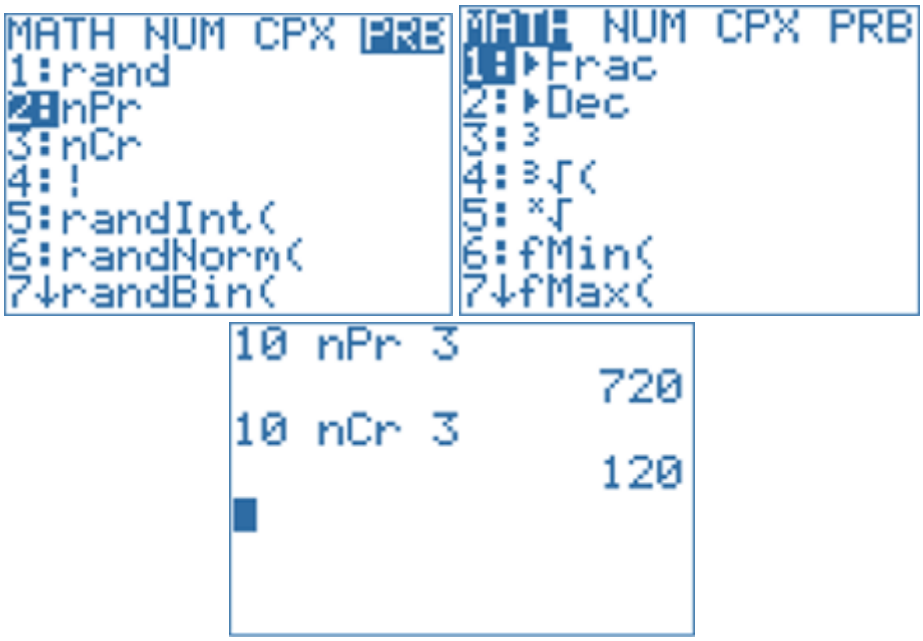
In practice, this means carefully identifying whether a problem involves arranging items in specific positions -- which calls for permutations -- or selecting items without regard to order -- which calls for combinations. Once that decision is made, technology becomes a powerful tool for handling large numbers, checking work, and exploring more complex applications. But the accuracy of every technological computation still depends on your ability to correctly interpret the structure of the problem.

Modern calculators include built-in commands for computing permutations and combinations, typically labeled as **nPr** and **nCr**. These functions automate the factorial calculations and eliminate arithmetic errors.

Press [MATH] and use the arrow to scroll to the right to the PRB (probability) menu.

You will see the choices for permutation and combination: 2:nPr, 3:nCr.

In both cases, you must type the value for n first, then select the command, then the value of r, and press ENTER. The screenshots below show the menu and the proper uses of these commands.



Using Multiple Counting Techniques

Some counting problems require more than one technique. In particular, you may need to compute combinations for one part of a problem and then use the *Fundamental Counting Principle* to combine several independent choices.

Consider the following situation: A sandwich shop offers a special in which you must choose one type of bread, one type of protein, and three toppings from the menu. We want to determine how many different ways the special can be ordered.

Sandwich Menu



This problem resembles earlier examples where we used the *Fundamental Counting Principle*, but with one important difference: in previous cases, we selected one item from each category. Here, we still choose one bread and one protein, but we must select three distinct toppings, which requires a combination rather than a single choice.

We apply two ideas together: combinations for the toppings and the *Fundamental Counting Principle* for the overall count.

- Bread: There are 2 types of bread, so there are 2 ways to choose bread.

- Protein: There are 4 types of protein, so there are 4 ways to choose protein.
- Toppings: There are 5 vegetables, and we must choose 3 of them. Since we are choosing a group of 3 vegetables, count combinations ${}_5C_3$.

By the *Fundamental Counting Principle,* the total number of different specials is:

$$2 \times 4 \times {}_5C_3 = 2 \times 4 \times 10 = 80.$$

So, there are 80 different ways to select menu items for the sandwich special.

Example 7.2.6

Brenda plans to rent 5 movies for the weekend: 3 science fiction and 2 comedies. There are 6 science fiction movies available and 4 comedies available. How many different groups of 5 movies can she choose?

Solution

This selection happens in two stages.

- Select 3 movies from 6 science fiction movies**.** This is a group selection so use we use a combination. On a calculator, find ${}_6C_3 = 20$.
- Select 2 movies from 4 comedies**.** This is another combination. On a calculator evaluate ${}_4C_2 = 6$.

Each possible group of science fiction movies can be paired with each possible group of comedies. By the *Fundamental Counting Principle,* multiply the results:

$${}_6C_3 \times {}_4C_2 = 20 \times 6 = 120.$$

Brenda can choose her group of 5 movies in 120 different ways.

Example 7.2.7

An art gallery owns 11 paintings by a particular artist: 5 oil paintings and 6 watercolor paintings. For an upcoming special exhibition, the gallery will display 7 paintings, consisting of 3 oil paintings and 4 watercolor paintings. How many different exhibitions can be planned?

Solution

We break the selection into two stages:

Select 3 of the 5 oil paintings. Use a combination. On a calculator, evaluate ${}_5C_3 = 10$.

Select 4 of the 6 watercolor paintings. Again, use a combination. On a calculator, evaluate ${}_6C_4 = 15$.

Each possible group of oil paintings can be paired with each possible group of watercolor paintings. By the *Fundamental Counting Principle,* multiply the results:

$${}_5C_3 \times {}_6C_4 = 10 \times 15 = 150.$$

There are 150 different exhibitions the gallery could create using the artist's 11 paintings.

You Try It 7.2.E

A music director is preparing a small concert program featuring 6 performers. The orchestra has 9 violinists and 7 cellists available. For this program, the director wants to include 4 violinists and 2 cellists.

How many different groups of performers can be selected?

Answer available at the end of this section.

Permutations with Identical Items

When we count permutations, we normally assume that all objects are distinct. That is, we can distinguish each object from another. In those cases, the number of permutations of n objects is simply $n!$. But many real situations involve objects that are not all different -- letters that repeat, identical tiles, matching cards, or any scenario where some items cannot be distinguished from one another. In these cases, treating every object as unique would overcount the number of arrangements.

Permutations with identical correct this overcounting by dividing out the arrangements that look identical because of repeated items. To illustrate, imagine arranging the letters in the word **MOM**. Here, the two M's are identical. If all three letters were distinct, there would be $3! = 6$ possible permutations.

But because the two M's are identical, many of those permutations look the same. For example. you could not distinguish between MMO and MMO. These two arrangements are different only if we pretend the M's are distinguishable. In reality, they are not, so they represent one unique arrangement.

To ensure that swapping identical M's does not count as a "new" arrangement, we divide the total number of permutations by the number of ways the repeated items (the two M's) could be rearranged among themselves $2!$:

$$\frac{3!}{2!} = \frac{6}{2} = 3$$

So there are 3 distinct permutations of MOM.

This thinking leads us to generalize how to count permutations when some items are not distinct.

Permutations of Identical Items

If a collection contains items that appear identical, the number of distinguishable permutations is

$$\frac{n!}{n_1!\, n_2!\, n_3! \cdots}$$

where n is the total number of items, and each n_i represents the number of times a particular item is repeated.

Example 7.2.8

Find the number of distinguishable arrangements using letters from the word *MISSISSIPPI*.

Solution

Because several letters repeat, many permutations of these 11 letters would look identical if we treated each letter as distinct. To avoid overcounting, we use the formula for permutations of identical items.

The word contains 11 letters, so $n = 11$.

The repeated letters are:

- **I** appears 4 times, so $n_1 = 4$.
- **S** appears 4 times, so $n_2 = 4$.
- **P** appears 2 times so $n_3 = 2$.

Using the formula for permutations with non-distinct objects:

$$\begin{aligned}\frac{n!}{n_1! \cdot n_2! \cdot n_3!} &= \frac{11!}{4! \cdot 4! \cdot 2!} \\ &= \frac{39,916,800}{1,152} \\ &= 34,650\end{aligned}$$

There are 34,650 distinguishable arrangements of the letters in MISSISSIPPI.

You Try It $7.2.F$

Find the number of distinguishable arrangements using letters from the word BALLOON.

Answer available at the end of this section.

Example $7.2.9$

Jerry bought 5 red light bulbs and 3 blue light bulbs to place in a string of outdoor lights that has 8 sockets. How many different color patterns can Jerry create?

Solution

This situation is just like arranging letters with repetitions. We are forming an ordered pattern, but some bulbs are identical in color. If we treated all bulbs as distinct, we would overcount patterns that look the same.

There are 8 bulbs total, so $n = 8$. The repeated colors are red bulbs $(n_1 = 5)$ and blue bulbs $(n_2 = 3)$.

Use the formula for permutations with repeated objects:

$$\begin{aligned}\frac{n!}{n_1! \cdot n_2!} &= \frac{8!}{5! \cdot 3!} \\ &= \frac{40,320}{720} \\ &= 56\end{aligned}$$

Jerry can make 56 different color patterns with his bulbs.

You Try It $7.2.G$

A teacher is organizing a display shelf with 10 books. She has 6 copies of the same math textbook and 4 copies of the same science workbook. Because the copies within each subject are identical, many arrangements would look the same if we treated every book as distinct. How many different book arrangements can she create on the shelf?

Answer available at the end of this section.

Section Summary

Counting problems often involve forming arrangements or groups from a set of objects. The method you choose depends on whether order matters and whether objects are distinct.

Permutations count ordered arrangements. Permutations are used whenever the order of the selected objects is important. When

selecting and arranging r objects from n distinct objects, the number of permutations is

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

Combinations count unordered groups. Combinations are used whenever the group itself matters but the arrangement within the group does not. When selecting r objects from n distinct objects, the number of combinations is

$$ {}_nC_r = \frac{n!}{(n-r)!\ r!}. $$

Many real problems require more than one technique. For example, you may form a group using combinations and then use the *Fundamental Counting Principle* to combine several independent selections. Recognizing when to blend techniques is an essential skill in solving multi-stage counting problems.

Finally, when objects are not distinct, treating them as distinct would overcount arrangements that look identical. To correct this, divide out the permutations of the repeated items. If a collection of n objects contains repeated groups of sizes n_1, n_2, n_3 ,..., the number of distinguishable permutations is

$$ \frac{n!}{n_1!\, n_2!\, n_3! \cdots}. $$

This formula ensures that swapping identical objects does not create a new arrangement.

Section 7.2 Review Questions

1. Simplify the following by hand:
   a. $6!$
   b. ${}_7P_3$
   c. $\dfrac{12!}{(12-3)!3!}$
   d. $\dfrac{10!}{7!}$
   e. $\dfrac{8!}{3!2!}$
2. Eight actresses are available to cast in a play that has 3 female roles. In how many ways could the 3 roles be cast?
3. You just bought 3 extra tickets to the Nationals game and you have 5 friends you can invite. How many ways can you select the group of friends to join you at the game?
4. How many different 5-card poker hands can be dealt from a 52-card deck?
5. In how many ways can first, second, and third prizes be awarded in a contest with 210 contestants?
6. A child wants to pick three pieces candy to take in her school lunch box. If there are 13 pieces of candy to choose from, how many ways can she pick the three pieces?
7. A college class has a reading list of eight books. A student must choose and read five of the books before the end of the course. In how many ways can the student choose five books to read?
8. A pianist plans to play 4 pieces at a recital. In how many ways can she arrange these pieces in the program?
9. The Volunteer Club has 18 members. A committee of members will be selected to plan the annual food drive. How many different 3-person committees can be selected?
10. A jury pool consists of 20 people. How many different 12-person juries can be selected from the jury pool?
11. A swimming event has 16 contestants. The swimmers with the five fastest speeds will be listed, in the order of their speed, on the leader board. How many ways are there for the names of five swimmers to be listed?

12. At a baby shower 17 guests are in attendance and 5 of them are randomly selected to receive a door prize. If all 5 prizes are identical, in how many ways can the prizes be awarded?

13. There are 12 standby passengers who hope to get on a flight to Hawaii, but only 6 seats are available on the plane. How many different ways can the 6 people be selected?

14. In how many ways can an investor put together a portfolio of five stocks and six bonds selected from her favorite nine stocks and eight bonds?

15. There are 4 Democrats and 6 Republicans on a Senate committee. How many different subcommittees can be formed with 1 Democrat and 2 Republicans?

16. To promote reading, a teacher decides to feature 3 classics, 4 contemporary novels, and 2 non-fiction books on a bulletin board. How many selections can she make from 5 classics, 6 contemporary novels and 4 non-fiction books?

17. Determine the number of distinguishable arrangements you make using the letters in each word:
    a. FACTOR
    b. MAMMAL
    c. STATISTICS

18. A ship's captain sends signals by arranging 4 yellow, 3 blue, and 1 red flags on a vertical pole. How many different signals could the ship's captain possibly send?

Answers

You Try It $7.2.A$
360,360

You Try It $7.2.B$
5,040

You Try It $7.2.C$
165

You Try It $7.2.D$
364

You Try It $7.2.E$
2,646

You Try It $7.2.F$
1,260

You Try It $7.2.G$
210

This page titled 7.2: Permutations and Combinations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.3: Probability Using Counting Techniques

Learning Objectives

- Use the *Fundamental Counting Principle*, permutations, and combinations to find probability.

Introduction

To compute theoretical probability using the formula \(P(E)=\frac{n(E)}{n(S)\), we must know how many outcomes are in the sample space S and how many outcomes belong to the event E. When these numbers are small, we can list and count the outcomes directly. However, many real-world situations involve far too many outcomes to list, so we rely on counting techniques to determine these quantities efficiently.

In this section, you will apply what you have learned about permutations, combinations, and the *Fundamental Counting Principle* to count the number of possible outcomes in a sample space or event. Once these counts are known, we will use them in the probability formula to compute theoretical probabilities accurately.

Probability with the *Fundamental Counting Principle*

In finding any probability, we follow the same three steps:

- First, count the outcomes in the sample space, $n(S)$.
- Next, count the favorable outcomes in the event, $n(E)$.
- Finally, substitute these counts into the probability formula, $P(E) = \dfrac{n(E)}{n(S)}$.

The next few examples demonstrate how the *Fundamental Counting Principle* allows us to quickly determine $n(S)$ and $n(E)$.

Example 7.3.1

Each question on a 5-question multiple-choice quiz has four possible answer choices: A, B, C, and D. Randy randomly guesses on every question.

What is the probability that he gets all a perfect score (all five questions correct)?

Solution

First, count the total number of possible answer patterns.

> Each question has 4 possible choices, and there are 5 questions. By the *Fundamental Counting Principle* the total number of possible answer patterns is $4 \times 4 \times 4 \times 4 \times 4 = 4^5 = 1024$.
>
> There are 1,024 different ways Randy could answer the quiz. This corresponds to the sample space so $n(S) = 1,024$.

Next, count the number of favorable answer patterns for the event "all 5 questions correct."

> Randy must choose the one correct answer for each of the 5 questions. There is exactly 1 way to do this: the unique pattern where every answer matches the key.
>
> There is 1 way that Randy could answer all the questions correctly for a perfect score. This corresponds to the event so $n(E) = 1$.

Finally, calculate the probability.

$$P(E) = \frac{n(E)}{n(S)} = \frac{1}{1024}.$$

So the chance that Randy gets a perfect score by random guessing is $\frac{1}{1024} \approx 0.00098$ or about 0.098%.

Example 7.3.2

The letters in the word **HOLIDAY** are randomly scrambled. What is the probability that the first two letters in the arrangement are both vowels?

Solution

First, count the number of ways that the 7 letters could be scrambled. Note that that the letters in **HOLIDAY** are distinct.

By the *Fundamental Counting Principle,* the number of possible permutations is $7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5,040$. This corresponds to the sample space so $n(S) = 5,040$.

Next, count the arrangements where the first two letters are vowels.

- The vowels in **HOLIDAY** are O, I, A -- three vowels total.
- There are 3 choices for the first letter (any vowel).
- After choosing one vowel, 2 choices remain for the second letter.
- Once the first two positions are filled with vowels, the 5 remaining letters can be arranged in $5!$ ways.

Using the *Fundamental Counting Principle,* there are $3 \cdot 2 \cdot 5! = 720$ favorable arrangements for this event. So, $n(E) = 720$.

Finally, calculate the probability.

$$P(E) = \frac{n(E)}{n(S)} = \frac{720}{5040} = \frac{1}{7}.$$

The probability that the first two letters in the arrangement are both vowels is $\frac{1}{7} \approx 0.1429 \text{ or about } 14.29\%$.

You Try It 7.3.A

A locker combination consists of 7 distinct digits arranged in a random order. The digits available are 0, 1, 2, 3, 4, 5, 6. What is the probability that the first three digits in the combination are even numbers?

Answer available at the end of this section.

Example 7.3.3

Banks often assign a 4-digit PIN (personal identification number) for debit cards. Assume these PINs are generated at random. What is the probability that a randomly assigned PIN has no repeated digits?

Solution

First, count all possible 4-digit PINs. Each digit can be any of the ten digits 0–9.

By the *Fundamental Counting Principle,* the number of PINS possible is $10 \cdot 10 \cdot 10 \cdot 10 = 10,000$. There are 10,000 possible PINS (from 0000 to 9999). This corresponds to the sample space so $n(S) = 10,000$.

Next, count the PINs with no repeated digits.

To form a PIN so that no digit is repeated, there are 10 choices for the first digit, 9 choices for the second digit, 8 choices for the third digit, and 7 choices for the fourth digit.

Using the *Fundamental Counting Principle,* there are $10 \cdot 9 \cdot 8 \cdot 7 = 5,040$ favorable arrangements for this event: $n(E) = 5,040$.

Finally, calculate the probability.

$$P(\text{no repeated digits}) = \frac{n(E)}{n(S)} = \frac{5,040}{10,000} = 0.504 = 50.4\%.$$

You Try It 7.3.B

A security system generates a 3-character access code using the letters A–Z (uppercase only). Each character is chosen at random.

What is the probability that the access code contains no repeated letters?

Answer available at the end of this section.

Probability with Permutations and Combinations

As we continue exploring probability, we now turn to situations where the order of selection matters or where we are forming groups without repetition. In these cases, the *Fundamental Counting Principle* still underlies our reasoning, but the counting becomes more structured. This is where permutations and combinations become essential tools.

Permutations and combinations apply only when the objects being arranged or selected cannot be repeated. If changing the order produces a different outcome, we use a permutation. If the order does not affect the outcome, we use a combination. The next set of examples will show how these formulas allow us to compute probabilities efficiently when order and non-repetition play a central role.

In these problems, the same three steps will be used: Find $n(S)$, next $n(E)$, and finally use them to calculate $P(E)$.

Example 7.3.4

A word game requires players to select 4 tiles from a bag containing 26 tiles, each labeled with one of the letters A through Z.

What is the probability that a player will be able to spell CATS with the tiles he selects?

Solution

First, count the number of possible groups of any 4 tiles that could be selected from the bag of 26 tiles.

Since the order of the tiles does not matter when selecting a group of tiles, we evaluate the combination ${}_{26}C_4 = 14,950$.

There are 14,950 ways to select a group of any 4 tiles. This corresponds to the sample space so $n(S) = 14,950$.

Next, count the favorable outcomes.

To spell **CATS**, the player must select exactly the tiles C, A, T, S but the player can do so in any order. There is only one one way this can happen so $n(E) = 1$.

If you are not convinced this is indeed only 1 way, we are choosing 4 tiles and they must be selected only from the tiles A, C, S, T:

$$_4C_4 = 1$$

Finally, calculate the probability.

$$P(E)) = \frac{n(E)}{n(S)} = \frac{1}{14,950}.$$

This probability that a player can spell CATS with the tiles is $\frac{1}{14,950} \approx 0.000067 \text{ or } 0.0067\%$. This is a very small likelihood!

Example 7.3.5

Professor Jones's desk contains 20 whiteboard markers, and 12 of them are depleted. He randomly selects five markers for class.

What is the probability that all five markers selected are depleted?

Solution

First, count the number of possible groups of any 5 markers that Professor Jones could select from the 20 markers in his desk.

Because we are counting groups of markers, we evaluate the combination $_{20}C_5 = 15,504$.

There are 15,504 ways to select a group of any 5 markers. This corresponds to the sample space so $n(S) = 15,504$.

Next, count the favorable outcomes.

This event contains the outcomes where all 5 markers are depleted. They must be chosen from the 12 depleted markers in the desk. To count the number of ways this can occur evaluate $_{12}C_5 = 792$.

Finally, calculate the probability.

$$P(E) = \frac{n(E)}{n(S)} = \frac{792}{15,504} = \frac{33}{646}$$

The probability that all 5 markers are depleted is $\frac{33}{646} \approx 0.0511 \text{ or } 5.11\%$.

You Try It 7.3.C

A drawer contains 10 batteries, and 6 of them are fully charged. Three batteries are randomly selected for a device.

What is the probability that all three batteries selected are fully charged?

Answer available at the end of this section.

Example 7.3.7

A school hosts an academic challenge featuring 8 competitors, and 4 of them are freshmen. All competitors have an equal chance of finishing in any position.

What is the probability that freshmen finish in 1st, 2nd, and 3rd place?

Solution

Because order matters in determining who finishes 1st, 2nd, and 3rd, this is a permutation probability problem.

First, count the number of ways 1st, 2nd, and 3rd place can be awarded among the 8 competitors.

Evaluate ${}_8P_3 = 336$. This corresponds to the sample space so $n(S) = 336$.

Next, count the favorable outcomes.

Count the number of ways 1st, 2nd, and 3rd place can be awarded among the 4 freshmen.

Evaluate ${}_4C_3 = 24$ so $n(E) = 24$.

Finally, calculate the probability.

$$P(E)) = \frac{n(E)}{n(S)} = \frac{24}{336} = \frac{1}{14}.$$

The probability that 1st, 2nd, and 3rd places are awarded to freshmen is $\frac{1}{14} \approx 0.0714 \text{ or } 7.14\%$.

It should be noted here that probability problems involving a permutation can often be solved in more than one way if you think flexibly. In Example 7.3.7, we can also apply the *Multiplicative Rule of Probability* for dependent events. Suppose that F_1 is the event that 1st place goes to a freshmen, F_2 is the event that 2nd place goes to a freshmen, and F_3 is the event that 3rd place goes to a freshmen.

$$\begin{aligned} P(F_1 \cap F_2 \cap F_3) &= P(F_1) \cdot P(F_2 \mid F_1) \cdot P(F_3 \mid F_1 \cap F_2) \\ &= \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{2}{6} \\ &= \frac{24}{336} \\ &= \frac{1}{14} \end{aligned}$$

This matches the result obtained using permutations.

Example 7.3.8

A bag contains 10 colored marbles: 6 red and 4 yellow. You randomly select 5 marbles.

What is the probability that all 5 marbles selected are red?

Solution

Because order does not matter, we use combinations to count the number of possible groups of 5 marbles.

First, count the number of ways of selecting 5 marbles from the bag.

Evaluate ${}_{10}C_5 = 252$. This corresponds to the sample space so $n(S) = 252$.

Next, count the favorable outcomes.

Count the number of ways to select 5 red marbles: $n(E) =_6 C_5 = 6$.

Finally, calculate the probability.

$$P(E) = \frac{n(E)}{n(S)} = \frac{6}{252} = \frac{3}{126}$$

The probability that all five marbles are red is $\frac{3}{126} = \approx 0.0238$ or 2.38%.

You Try It 7.3.D

A box contains 12 colored pencils: 7 blue and 5 green. You randomly select 4 pencils. What is the probability that all 4 pencils selected are blue?

Answer available at the end of this section.

Example 7.3.9

A bag contains 10 colored marbles: 6 red and 4 yellow. You randomly choose 5 marbles.

What is the probability that exactly 3 of them are red?

Solution

The sample space is the same as in Example 7.3.8. There are $n(S) = 252$ ways of selecting 5 marbles from the bag.

To count the number of favorable outcomes in this event, we need to select 5 marbles so that 3 are red. This means the other 2 must be yellow.

Use the *Fundamental Counting Principle* and find the product of two combinations.

$$\begin{aligned} {}_6C_3 \times {}_4C_2 &= 20 \times 6 \\ &= 120 \end{aligned}$$

So, $n(E) = 120$.

Finally, calculate the probability.

$$P(E) = \frac{n(E)}{n(S)} = \frac{120}{252} = \frac{10}{21}$$

The probability that exactly 3 marbles are red is $\frac{10}{21} \approx 0.4762$ or 47.62%.

Example 7.3.10

Calculate the probability of being dealt exactly two hearts in a five-card poker hand.

Solution

A standard deck has 52 cards, including 13 hearts and 39 non-hearts. Because order does not matter in a poker hand, we use combinations.

First, count the number of ways of selecting 5 cards from the deck.

Eveluate ${}_5C_52 = 2,598,960$. This corresponds to the sample space so $n(S) = 2,598,960$.

Next, count the favorable outcomes.

Count the number of ways to select 2 hearts. But remember this means the other 3 cards are not hearts.

Use the *Fundamental Counting Principle* and find the product of two combinations.

$$\begin{aligned} {}_{13}C_2 \times {}_{39}C_3 &= 78 \times 9,139 \\ &= 712,542 \end{aligned}$$

So, $n(E) = 712,542$.

Finally, calculate the probability.

$$P(E) = \frac{n(E)}{n(S)} = \frac{712,542}{2,598,960}.$$

The probability of selecting 2 hearts in a five-card poker hand is $\frac{712,542}{2,598,960} \approx 0.2743$ or 27.43%.

You Try It $7.3.E$

Calculate the probability of being dealt four aces in a five-card poker hand.

Answer available at the end of this section.

Section Summary

Counting techniques allow us to analyze a wide variety of probability problems. In this section we used *Fundamental Counting Principle,* permutations, or combinations to count outcomes so we could apply the formula for theoretical probability. The choice depends on whether the situation involves sequential choices, ordered arrangements, or unordered selections.

To find theoretical probability efficiently and effectively, every problem follows the same three steps:

1. Count the outcomes in the sample space, $n(S)$.
2. Identify which outcomes satisfy the event and count the favorable outcomes in the event, $n(E)$.
3. Substitute these counts into the probability formula $P(E) = \frac{n(E)}{n(S)}$.

Section 7.3 Review Questions

1. Four students with different last names randomly line up to go to lunch. What is the probability that they are lined up in alphabetical order by last name?

2. What is the probability that a randomly assigned 5-digit student ID number has all odd digits? Assume digits may be repeated.

3. The letters in the word NUMBER are arranged in random order. What is the probability that the first two letters are both consonants?

4. A high school is staging the musical *West Side Story*. There are 20 singers auditioning for the musical. (We will assume they are equally talented.)
   a. The director is looking for two singers who could sing a good duet. What is the probability that Alicia and Juan are the two singers who are selected by the director?
   b. The director is also interested in the number of ways to choose a lead singer and a backup singer. What is the probability that Alicia is selected the lead singer and Juan is selected the backup singer?

5. A motorcycle license plate consists of 5 digits that are randomly selected. No digit is repeated. What is the probability of getting a license plate with all odd numbers?

6. A technician is launching fireworks near the end of a show. Of the remaining 15 fireworks, 8 are blue, 4 are red, and 3 are white. If she launches 5 of them in a random order, what is the probability that she launches all blue fireworks?

7. There are 12 males and 8 females in a jury pool. A group of 8 of them will be selected to sit on a jury. What's the probability of selecting 5 males and 3 females for the jury?

8. There are 10 canned drinks in a cooler: 6 are colas and 4 are fruit drinks. Danielle will reach into the cooler and pull out 5 drinks without looking. Find the probability that
   a. all 5 drinks are colas.
   b. 3 drinks are colas and 2 drinks are fruit drinks.

9. Compute the probability that a 5-card poker hand is dealt to you that contains
   a. all diamonds
   b. two Queens.

10. A bag contains 6 real diamonds and 5 fake diamonds. If 4 diamonds are picked from the bag at random, what is the probability that
    a. none of them are fake.
    b. exactly 2 of them are real.

Answers

You Try It $7.3.A$

$$\frac{576}{5{,}040} = \frac{4}{35} \approx 0.1143 \approx 11.43\%$$

You Try It $7.3.B$

$$\frac{15{,}600}{17{,}576} = \frac{150}{169} = 0.8875 = 88.75\%$$

You Try It $7.3.C$

$\frac{20}{120} = \frac{1}{6} \approx 0.1667$ or 16.67%

You Try It $7.3.D$

$\frac{35}{495} = \frac{7}{99} \approx 0.0707$ or 7.07%.

You Try It $7.3.E$

$\frac{48}{2{,}598{,}960} \approx 0.000018$ or 0.0018%

This page titled 7.3: Probability Using Counting Techniques is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

8: APPENDIX- ANSWERS AND FREE GRAPHING TOOLS

8.1: Answers to Section Review Exercises
8.2: Free Online Graphing Tools

This page titled 8: Appendix- Answers and Free Graphing Tools is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

8.1: Answers to Section Review Exercises

Answers to Section Review Exercises

Chapter 1

Section 1.1

1. a. not statistical
   b. not statistical
   c. not statistical
   d. statistical
   e. statistical
   f. statistical
   g. not statistical
   h. not statistical
   i. statistical
   j. not statistical
   k. statistical
2. a. categorical
   b. quantitative
   c. categorical
   d. quantitative
   e. categorical
   f. categorical
   g. quantitative
   h. quantitative
   i. categorical
3. a. continuous
   b. discrete
   c. continuous
   d. continuous
   e. discrete
   f. discrete
   g. discrete

Section 1.2

1. a. *individual:* a chocolate bar made by Sweet Tooth Candy Company; *variable:* weight of a chocolate bar (quantitative); *population:* all the chocolate bars made by Sweet Tooth Candy Company; *sample:* every 100th candy bar selected to be weighed.
   b. *individual:* an entree that Lucio serves in his restaurant; *variable:* temperature of an entree (quantitative); *population:* all entrees that Lucio serves in his restaurant; *sample:* 10 entrees that Lucio for which temperature was measured
   c. *individual:* a Maryland worker aged 18 or older; *variable:* mode of transportation used to get to work (categorical); *population:* all Maryland workers aged 18 or older; *sample:* 500 workers selected for the survey
   d. *individual:* a heart attack patient; *variable:* cholesterol level two days after having the heart attack (quantitative); *population:* heart attack patients two days after having a heart attack;*sample:* the 8 heart patients about whom data were collected
2. a. statistic
   b. parameter
   c. parameter
   d. parameter

e. statistic
f. parameter

Section 1.3

1. a. ratio
   b. nominal
   c. ordinal
   d. interval
   e. ordinal
   f. ratio
2. a. nominal
   b. ordinal
   c. interval
   d. ratio
   e. ordinal
   f. nominal

Section 1.4

1. a. Answers may vary, but boys and players who go to other schools will be underrepresented.
   2. selection bias. The way the sample is selected may create bias.
   3. sampling frame
   4. stratified random sample
2. response bias. You may be sensitive to the fact that a snake is asking you these questions. Your responses to whether or not you own items made of snakeskin may be different because a snake is asking you the questions. The relationship between the interviewer and the interviewee is introducing bias in the data.
3. a. response bias (They may not recall or may overinflate their income.)
   b. non-response bias
   c. response bias (The question is leading/biased.)
   d. selection bias (The sample is collected based on ease.)
   e. response bias (Students may be embarrassed or fear stigma or consequences.)
4. a. cluster sample
   b. systematic sample
   c. convenience sample
   d. simple random sample
   e. stratified sample
5. a. simple random sample
   b. convenience sample
   c. stratified sample
   d. systematic sample
   e. cluster sample
6. a. convenience sample (The first 30 riders may be most excited to ride it since they were first in line.)
   b. cluster sample (no obvious bias identified)
   c. voluntary response (Riders who have strongest feelings will be more likely to respond.)
   d. simple random sample (no obvious bias identified)
   e. systematic (The question is a leading question.)
   f. stratified (no obvious bias identified)
7. {2, 7, 12, 17, 22, 27, 32}

Chapter 2

Section 2.1

1. a. Once leveled out, each of the four stacks will contain 5 blocks. The mean is 5.
   b. Each of the six stacks will contain 3 whole blocks, but there are 3 blocks that remain. Each of these remaining blocks

can be cut into half and distributed to the six stacks. Once leveled, each of the six stacks will contain $3\frac{1}{2}$ blocks so the mean is $3\frac{1}{2}$.

2. a. at 9
   b. Answers can vary: at 8 and 10, at 7 and 11, at 6 and 12, or at 5 and 13.
   c. at 10
   d. at 12 and 13
3. mean is 31 and median is 32
4. mean is 21.75, median is 15, mode is 11. The median is most likely the best measure of center. The mean is sensitive to the extreme value of 165 while the median is resistant to it.
5. mode = 160, mean ≈ 158.6, median = 159
6. The mean height is greater than the median height because Lucas's height will influence the mean, but not the median.
7. Enrique must score 97 in the 4th quarter so his mean will be 92.5. This would round to be 93 and exempt him from the final.
8. 90
9. 7 hours
10. Answers will vary, but the sum of the scores must be 574 and the 4th score in order from smallest to largest must be 80. One example is $\{70, 74, 75, 80, 85, 90, 100\}$.
11. Answers will vary, but the sum of the ten rolls must be 30. The 5th and 6th roll in order from smallest to largest must add to be 7 (or average to be 3.5). On example is $\{1, 1, 1, 2, 3, 4, 4, 4, 5, 5\}$.
12. Answers will vary, but the sum of the ages must be 66. The 4th and 5th ages in order from smallest to largest must add to be 16 but cannot both be 8 because there is not mode. One example is $\{3, 4, 5, 7, 9, 10, 12, 16\}$.
13. about 123.33 pounds
14. $691
15. $5.79
16. 83.7
17. a. 2.93
    b. Sharla's GPA will be higher than 2.93 because she increased her grade in a class worth 3 credits and lowered her grade in the class worth only 1 credit.
18. a. 82.05
    b. 76.8125
    c. at least 92.75

Section 2.2

1. a. 82; 18
   b. 198
   c. 288
2. About 85% of the other customers waiting in line at the DMV have waited the same or less time as Mina. Only 15% have waited longer. This is probably not a good thing from Mina's perspective. She is waiting longer than most of the others who are waiting.
3. a. 80%
   b. 30%
   c. higher
4. a. 33rd percentile
   b. 291
5. The five-number summary is [29, 37, 47, 61, 104]. The fence values are 1 and 97 so there the value 104 is an outlier.
6. The fence values are 6 and 22. The minimum 5 and the maximum 24 are both outliers. There may be others but we cannot say without seeing the data.
7. The five-number summary is [8.3, 35.9, 63.8, 80.6, 123.4]. The fence values are -31.15 and 147.65. There are no outliers in these data.
8. The five-number summary is [20, 24, 28, 32, 35]. The fence values are 12 and 44, so the value 10 is an outlier.

Section 2.3

1. a. 11 feet
   b. 6.5 feet
   c. 3.15625 feet

2. a. 69.3 inches
   b. ≈ 230.68 square inches and ≈ 15.19 inches
3. a. 20
   b. 10
   c. ≈ 5.14
   d. ≈ 44.33 and ≈ 6.66
   e. calculator
   f. 2
   g. 0
   h. ≈ 0.75
4. a. 1.5
   b. -2.25
   c. 0
5. ≈ 0.67
6. a. 2.5
   b. 2.09
   c. While he did well on both, Sampson's performance on the biology quiz was slightly better than his performance because the z-score in biology is further above 0 and therefore further above the mean of the students from his class.
7. A 55° day in January has a z-score of 1.9, while in July a 55° day has a z-score of -2.375. Because July's z-score has a larger absolute value (it is further from z = 0), a 55° day in July is more unusual.
8. a. The drum set has the lowest z-score (-1) so it costs the least relative to other drum sets.
   b. The guitar has the highest z-score (0.25) so it costs the most relative to other guitars.
9. a. 92.22 ≈ 92
   b. 48.54 ≈ 49

Section 2.4

1. The mean and median will both increase by $2,000. The IQR and the standard deviation will not change.
   b. The mean, median, IQR, and standard deviation will all get multiplied by 110%.
   c. The mean and standard deviation will both increase. The median and IQR will remain unchanged.
2. The mean increases to 9.05 pounds but the standard deviation remains 1.25 pounds.
3. The mean is 108 ounces and the standard deviation is 20 ounces.
4. The mean will increase. The median may or may increase depending on the values of the remaining 4 scores.
5. The mean will increase. The median may or may increase depending on the values of the remaining heights.
6. a. decreases
   b. decreases
   c. stays the same
   d. stays the same
   e. decreases
7. a. decreases
   b. increases
   c. stays the same
   d. stays the same
   e. increases
8. a. $211.69
   b. $132.42
   c. $206.00
   d. $337.00
   e. $401.00

Chapter 3

Section 3.1

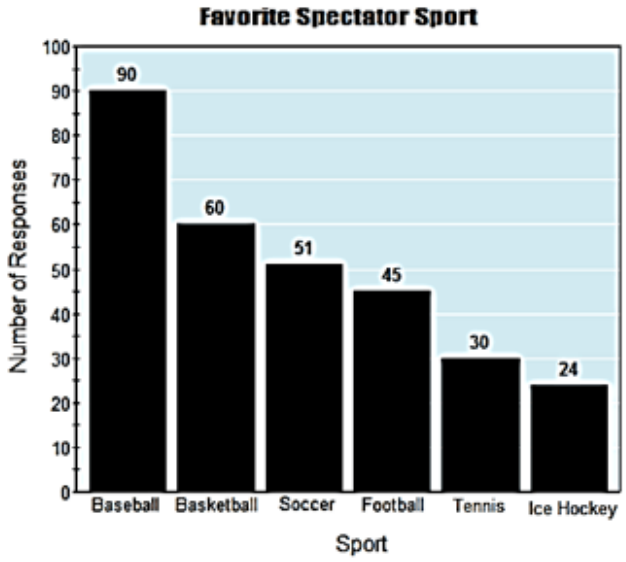
1. a. ordinal
   b. 19

c. ≈ 63%
d. 3

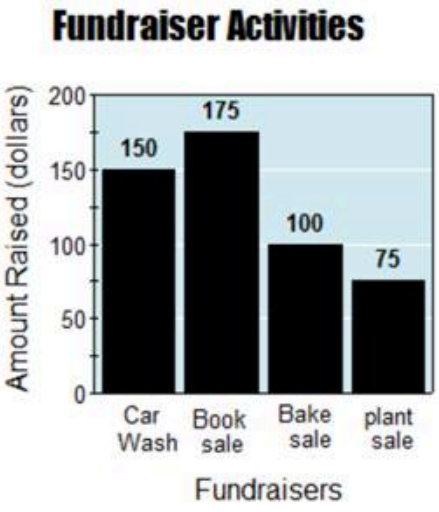
2. 5 students

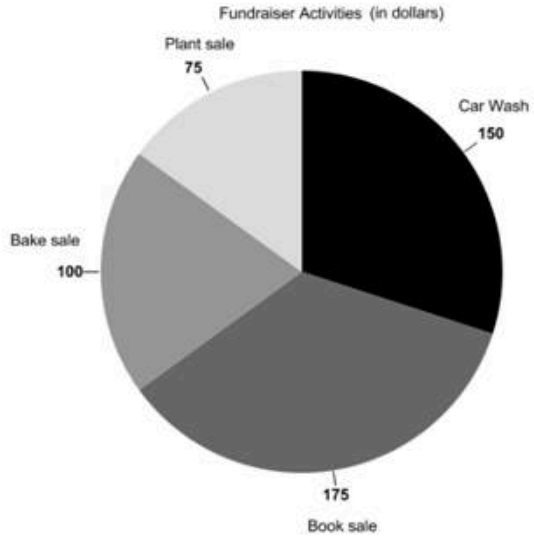
3. [redacted]

4. a. 15%
   b. 60
   c. 27



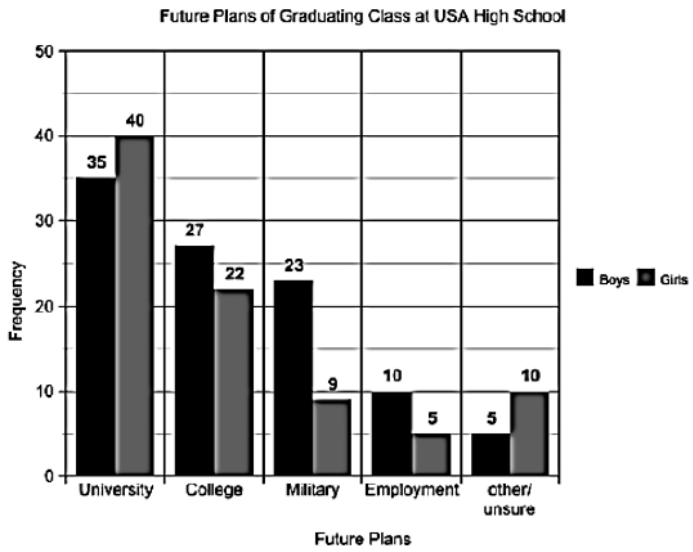
d.





5.

6. a. 5
   b. radio
   c. 45%



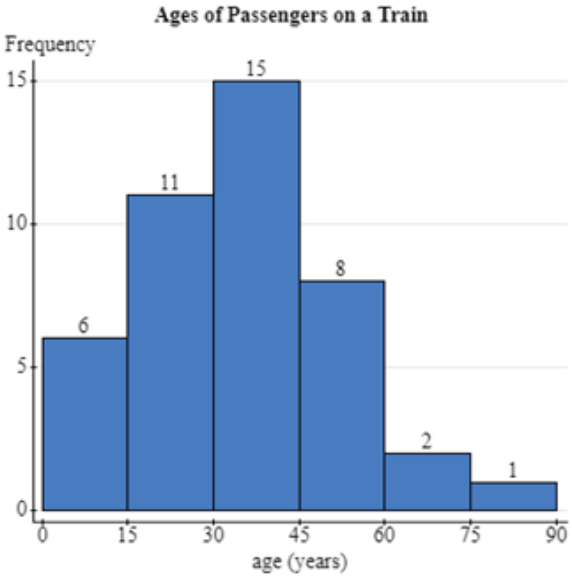
7.
8. a. 40
   b. 4
9. The pictogram is misleading because it violates the area principle and is misleading. The pictures are different sizes which may lead a reader to think more bananas were collected because the row of 4 bananas is longer than the row of 5 apples.
10. a.136, reading between the data
    b. ≈ 4.73 cars, reading beyond the data
    c. 36, reading the data
    d. ≈ 39%, reading between the data
    e. 20, reading between the data
    f. $5,000, reading beyond the data
11. Answers will vary. Here are some sample answers:
    a. How many employees have earned a high school/GED as their highest degree? 5
    b. What percent of the employees have earned a Master's degree? ≈ 19%
    c. How many employees have earned the highest degree possible? 22

Section 3.2

1. 
2.
3.

Ages of Passengers on a Train

4. 

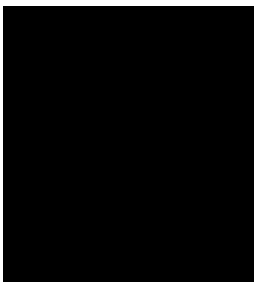


6.

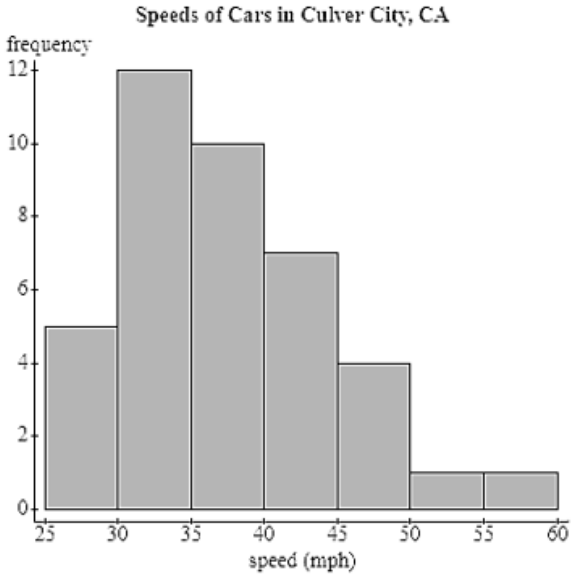
7. a. unimodal, symmetric
   b. bimodal, symmetric
   c. unimodal, skewed right
   d. unimodal, skewed left
8. a. unimodal, skewed right
   b. 50 mm
   c. 9 months
   d. 13/60 ≈ 22%
   e. [50, 100)
   f. Mean is larger than median because of right skewness
   g. about 250 mm using the midpoints of the first and last bins
9. symmetric and uniform. We expect about the same frequency for each of the results.
10. a. skewed right
    b. skewed left
    c. symmetric
11. Frequency plot. An ogive plot shows cumulative frequency and the segments could never fall. In each bin, there must be 0 or more data values included. Data cannot be removed.
12. a. 100
    b. 22
    c. 28
    d. 50
    e. [70, 80)
    f. [40, 50)

Section 3.3

1. a. unimodal, skewed left
   b. unimodal, skewed right
   c. bimodal, slightly skewed left
   d. unimodal, approximately symmetric with an outlier
2. a. skewed right and unimodal
   b. 2.6
   c. 1.5
   d. 1
   e. 9
   f. 3
   g. The fences are [-3.5, 8.5]. The student who read 9 books is an outlier.
3. a. III b. II c. I d. IV
4. a. I. unimodal, symmetric II. unimodal, approximately symmetric III. bimodal, approximately symmetric IV. uniform, symmetric
   b. As they are all symmetric or approximately symmetric, the mean would be the best measure of center.
   c. I
   d. III
5. a. 5 people, reading the data
   b. 12 minutes, reading the data
   c. 10 minutes, reading between the data
   d. $\frac{18}{40} = \frac{9}{20}$, reading between the data
   e. 2 people, reading beyond the data
6. a. 34
   b. 90
   c. 60
   d. 21
   e. ≈ 12%
   f. 8
   g. unimodal and skewed right
   h. mean
7. a. mean is 36.9 mph, median is 35.5 mph, mode is mode 32 mph

   b. 

c.



8.

Ages of Canadian Prime Ministers Upon Taking Office

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

3 | 9 = 39 years

The median is 54.5 years. The mode is 46 years.

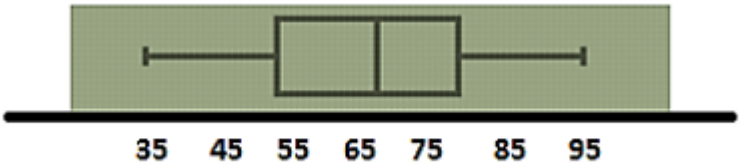
9. a. 70
   b. 4
   c. July
   d. none
   e. April
   f. June, August, September
   g. June and July
   h. commercial
   i. 60
   j. April
   k. 320
   l. The number of residential fires was the same in May as June.

Section 3.4

1. min = 4, Q1 = 8, median = 13, Q3 = 16, max = 20; Skewed to the left.
2. a. 75%
   b. 50%
   c. 25 guests

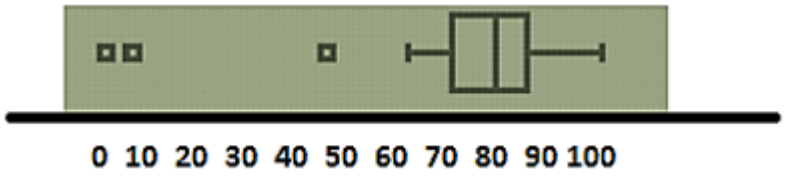
d. 34 guests

3. a. min = 35, Q1 = 53, median = 67, Q3 = 77.5, max = 95
   b. The fences are [-1.75, 114.25]. No data lie outside this interval. There are no outliers.



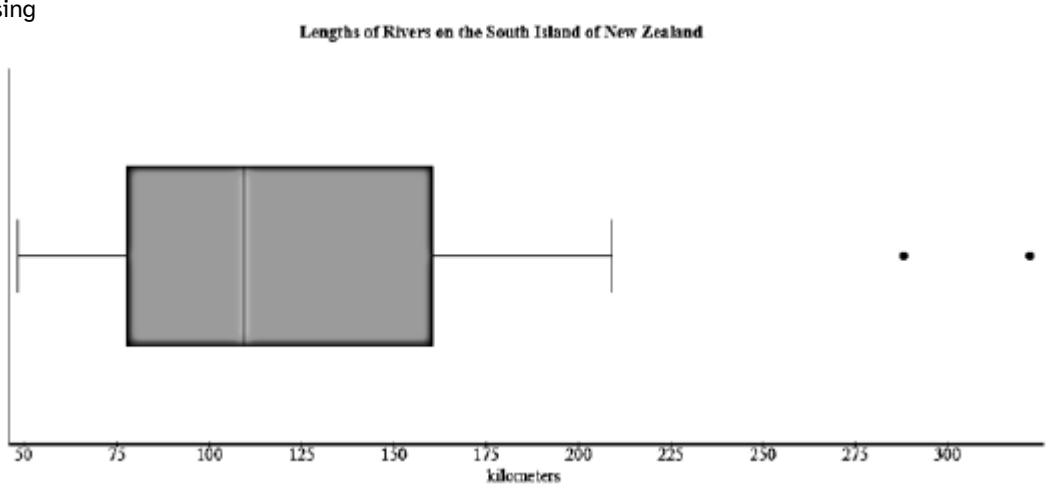
   c.
   d. approximately symmetric
   e. They should be roughly the same.
   f. median = 67 while mean ≈ 66.1

4. a. min = 0, Q1 = 72, median = 82, Q3 = 89, max = 105
   b. The fences are [46.5, 114.5]. There are 3 outliers: 0, 4, and 46.



   c.
   d. approximately symmetric with outliers on the left
   e. The mean should be smaller than the median because of the outliers on the left end.
   f. median = 82 while mean ≈ 75.4

5. Missing

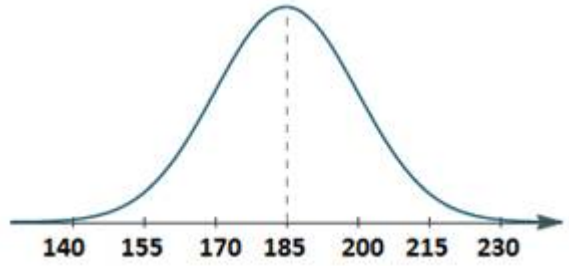
6. 

7. Data may vary. An example is 16, 17, 18, 19, 20, 20, 21, 22, 24, 25, 26.
8. f
9. a. 110 beats per minute
   b. after exercise
   c. before exercise
   d. after exercise
   e. after exercise
10. a. 75%
   b. 60 minutes
   c. It is more common for a teenager to spend more than 1 hour getting ready for a party. About 50% of students do this. Only about 25% spend more than 1 hour getting ready for school.
   d. i. True ii. False iii. True iv. True v. True

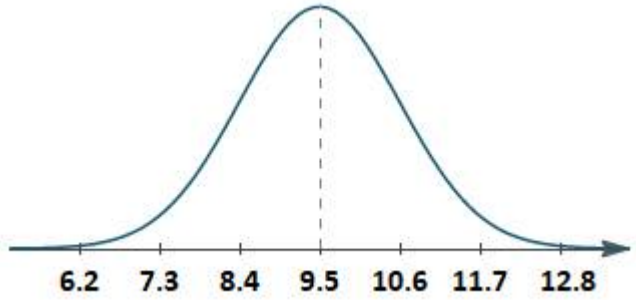
Section 3.5

1. $\mu = 15$; $\sigma = 3$

2. $\mu = -10$; $\sigma = 5$
3. The means of the 3 normal curves are different but the standard deviations are the same. The normal curve on the left has the smallest mean while the normal curve on the right has the largest mean.
4. The means of the 3 normal curves are all the same, but the standard deviations are different. The tallest normal curve has the smallest standard deviation while the shortest normal curve has the largest standard deviation.



5. a.
   b. i. 84% ii. 2.5% iii. 81.5%

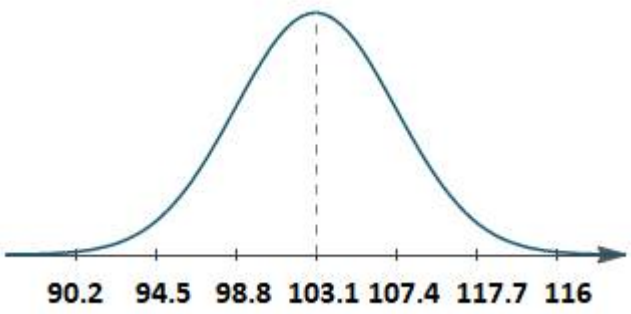


6. a.
   b. 16% ii) 95% iii) 0.15%
7. a. 81.5%
   b. 16%
8. a. The problem says we can assume dishwasher lifetimes are normally distributed.
   b. 81.5%
   c. 2.5%
9. The variable described in part a is most likely normally distributed. The variables in parts b and c are most likely skewed right. The variable in part d is most likely skewed left.
10. a. 2.56, 0, and 1.92
    b. 90
    c. 78
11. a. A score of 90 is more impressive in Class A. A score of 90 in Class A would have a z-score of 1.8 while a score of 90 in Class B would have a z-score of 1.6.
    b. A score of 60 is more discouraging in Class B. A score of 60 in Class A would have a z-score of -1.2 while a score of 60 in Class B would have a z-score of -4.4.
12. a. The Team 1 runner is faster. It took this runner less time than the runner from Team 2.
    b. The Team 2 runner is faster relative to his peers. The z-score of the runner from Team 1 is -1.71. The z-score of the runner from Team 2 is -3.42, so Team 2 runner's time is further below average (and faster) than Team 1 runner.

Section 3.6

1. a. ≈ 84%
   b. ≈ 41%
   c. ≈ 2%
2. a. ≈ 5%
   b. ≈ 16%
   c. ≈ 67%
   d. ≈ 0%
3. a. ≈ 71%
   b. ≈ 85%
   c. ≈ 97%

d. ≈ 46.0 cm



4. a.
   b. i. ≈ 33% ii. ≈ 3% iii. ≈ 95% iv. ≈ 64%
   c. ≈ 107.6 grams
5. a. ≈ 27%
   b. 5.79 cm and 5.91 cm
   c. 6.16 cm. About 90% of tangerines harvested on this farm have diameters of 6.16 cm or smaller. Only 10% of tangerines have larger diameters.
6. a. ≈ 37%
   b. 91st percentile
   c. ≈ 20
7. a. ≈ 0.0401
   b. ≈ 0.4938
   c. ≈ 2.25 ounces
8. a. ≈ 20
   b. 18.8 years
   c. ≈ 40.4 years

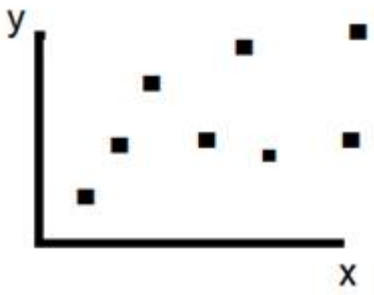
Chapter 4

Section 4.1

1. a. negative
   b. positive
   c. negative
   d. close to zero
   e. positive
   f. negative
   g. close to zero
2. Here are some samples:



a.



b.

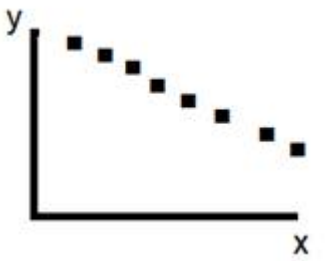
3. a. no correlation
   b. negative. positive
   c. positive
   d. negative
4. a. -0.05
   b. -0.78
   c. 0.75
   d. 0.92
   e. -0.95
5. a. The data will all lie on a straight line with positive slope:



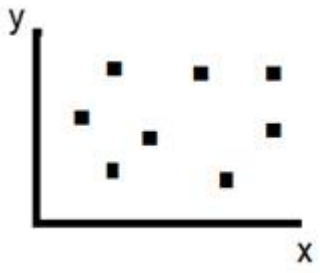
   b. The data will be slightly scattered along a straight line with negative slope:



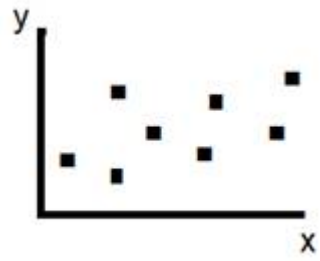
   c. The data will all lie on a straight line with negative slope:



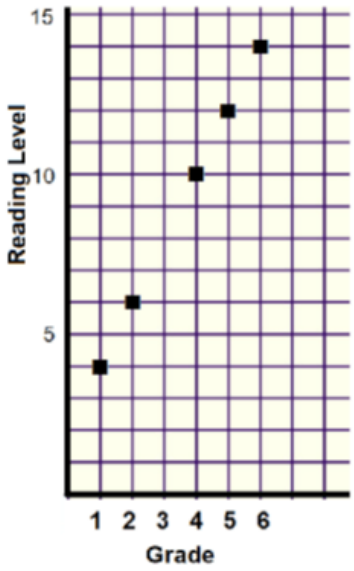
d. The data will be randomly scattered with no evidence of a linear trend:



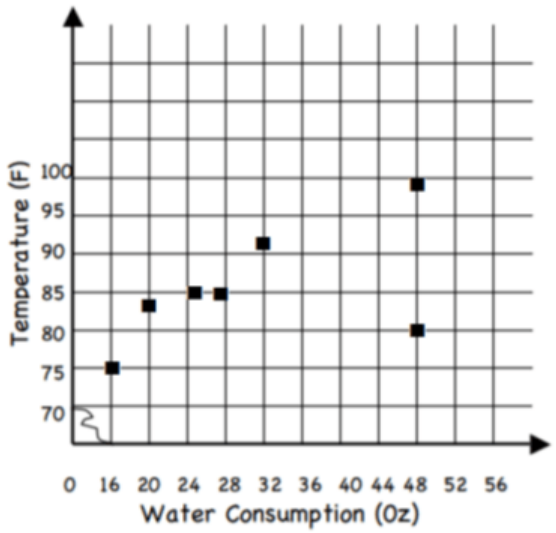
e. The data will be widely scattered along a straight line with positive slope:

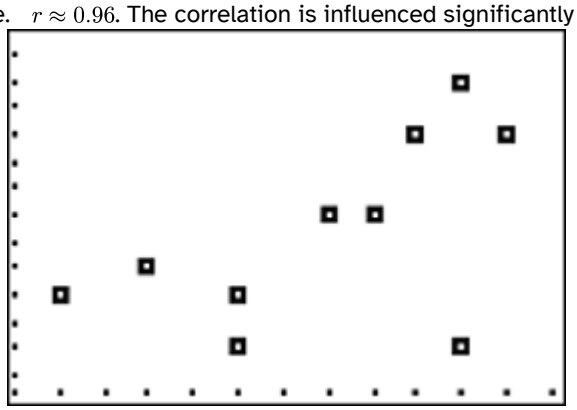


6. a.



b. The relationship is linear, positive, and strong without outliers.
c. $r = 1.0$



7. a.
   b. linear, positive, and strong relationship with one suspected outlier of 48 ounces of water and 80 degrees F.
   c. calculator
   d. $r \approx 0.55$
   e. $r \approx 0.96$. The correlation is influenced significantly by the outlier. The correlation is a non-resistant measure of strength.
8. a. 
   b. The relationship between the quiz scores is linear, positive, and moderately strong. Student #8 is a suspected outlier.
   c. $r \approx 0.57$

Section 4.2

1. a. study time
   b. amount of rain
   c. not possible
   d. weight of diamond
2. a. The explanatory variable is the number of years since 2001 and the response variable is the cost of tuition.
   b. $74.85
   c. The y-intercept is 51.75. The predicted cost of tuition in the year 2001 is $51.75.
   d. The slope is 1.65. Tuition is predicted to increase by $1.65 each year after 2001.
3. a. The explanatory variable is the number of miles on the car's odometer, and the response variable is the price of the car.
   b. $27,600
   c. The y-intercept is 28,400. The predicted price of a new car (a car not driven at all) is $28,400.
   d. The slope is-0.16. The price of a car is predicted to decrease by $0.16 (16 cents) for each additional mile it has been driven.
4. For $x = 1$, $y = 2$, $\hat{y} = 3.5$, and $e = -1.15$.

   For $x = 2$, $y = 8$, $\hat{y} = 4$, and $e = 4$.

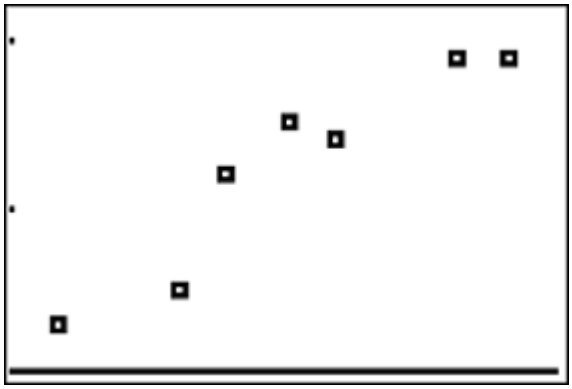
   For $x = 4$, $y = 3$, $\hat{y} = 5$, and $e = -2$.

For $x = 6$, $y = 7$, $\hat{y} = 6$, and $e = 1$.

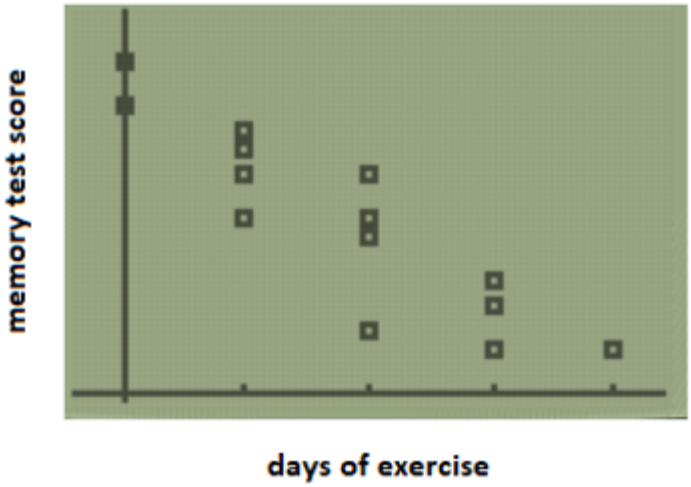
For $x = 8$, $y = 8$, $\hat{y} = 7$, and $e = 1$.

5. a. life expectancy
   b. The slope is 0.036. The life expectancy of an animal is expected to increase by 0.036 years for each additional day in its gestational period.
   c. overpredicts
   d. positive
   e. dog
   f. goat
   g. 3 years

Section 4.3

1. a. 
   b. Yes, there appears to be a strong positive correlation.
   c. $\hat{y} = 0.0974 + 0.0055x$
   d. $r \approx 0.9467$. The value is positive and its absolute value is close to 1.
   e. *Problem 1e*

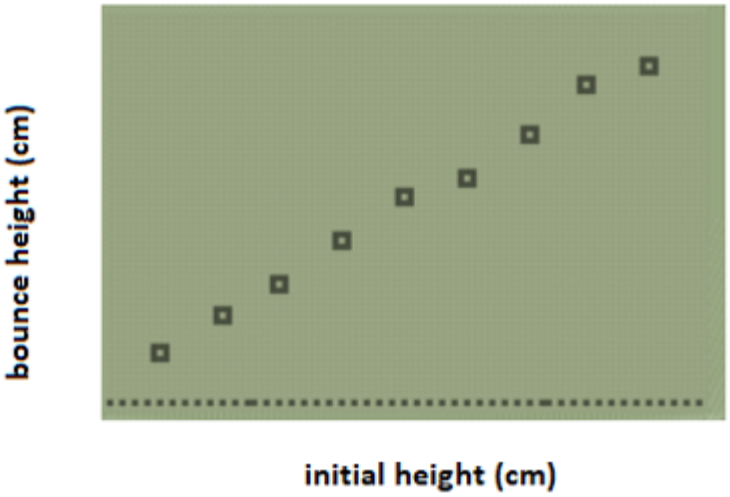
SAT math score, *x*	595	520	715	405	680	490	565
GPA, *y*	3.4	3.2	3.9	2.3	3.9	2.5	3.5
predicted GPA, $\hat{y}$	3.3699	2.9574	4.0299	2.3249	3.8374	2.7924	3.2049
residual, $y - \hat{y}$	0.0301	0.2426	-0.1299	-0.0249	0.0626	-0.2924	0.2951



2. a.
   b. The relationship appears to be linear because the data is clustered along a straight line.
   c. $\widehat{\text{test score}} = 13.2254 - 3.0915(\text{days})$
   d. 4 points
   e. The slope is -3.0915. We predict that a student's memory test score will decrease by about 3.0915 points for each additional day they exercise.
   f. The y-intercept is 13.2254. We predict a student's memory test score will be 13.2254 points if the students does not exercise at all.
   g. Student #2
   h. $1.1406 \approx 1$ point
   i. -8.4151. The prediction of a negative test score does not make sense. This was due to an extrapolation error of using a value for 'days' that was far outside the data set used to write this model.
3. a. Using [5, 52, 1] by [13, 50, 1], there appears to be a strong negative relationship between temperature and number of cups of coffee sold.



   b. No outliers are obvious.
   c. $r \approx -0.9503$
   d. $\widehat{\text{cups sold}} = 51.84 - 0.75(\text{temperature})$
   e. ≈ 35 cups
   f. It is possible to make this prediction, but it would not be appropriate to do so. A temperature of -10°F is far below the range of temperature values used to write the model so we might create extrapolation error in the prediction.
   g. The slope is -0.75. For every degree warmer day is, James and Keisha can expect to sell 0.75 fewer cups of coffee.
   h. ≈ 2 cups



4. a. The correlation is very strong because the data points almost lie perfectly on a straight line.
   b. $\widehat{\text{bounce height}} = -1 + 0.7667(\text{initial height})$
   c. The slope is 0.7667. We predict the bounce height will increase by 0.7667 cm when the initial height is increased by 1 cm.
   d. The y-intercept is -1 and has no meaning in this scenario.
   e. about 86 cm. There may be concerns because of extrapolation error. The highest bounce in the original data was only 52 cm and 65 cm is significantly larger. The same trend in the data may not continue past 52 cm.

Chapter 5

Section 5.1

1. a. $\frac{1}{12}$
   b. $\frac{3}{12} = \frac{1}{4}$
   c. $\frac{6}{12} = \frac{1}{2}$
   d. $\frac{6}{12} = \frac{1}{2}$
   e. 0
2. a. $\frac{7}{11}$
   b. $\frac{9}{11}$
   c. $\frac{2}{7}$
3. a. $\{(B1, B2), (B1, R), (B2, B1), (B2, R), (R, B1), (R, B2)\}$
   b. $\frac{2}{6} = \frac{1}{3}$
   c. $\frac{4}{6} = \frac{2}{3}$
   d. $\frac{2}{6} = \frac{1}{3}$
4. a. $\frac{4}{52} = \frac{1}{13}$
   b. $\frac{2}{52} = \frac{1}{26}$
   c. $\frac{12}{52} = \frac{3}{32}$
5. a. $\{(4, 2)(4, 6)(4, 8)(7, 2)(7, 6)(7, 8)(9, 2)(9, 6)(9, 8)\}$
   b. $\frac{6}{9} = \frac{2}{3}$
   c. $\frac{3}{9} = \frac{1}{3}$
   d. $\frac{5}{9}$
   e. $\frac{2}{9}$
6. a. $\{(A, E)(A, F)(B, E)(B, F)(C, E)(C, F)(D, E)(D, F)\}$
   b. $\frac{2}{8} = \frac{1}{4}$
   c. $\frac{3}{8}$
   d. $\frac{1}{8}$
   e. $\frac{4}{8} = \frac{1}{2}$
   f. $\frac{7}{8}$

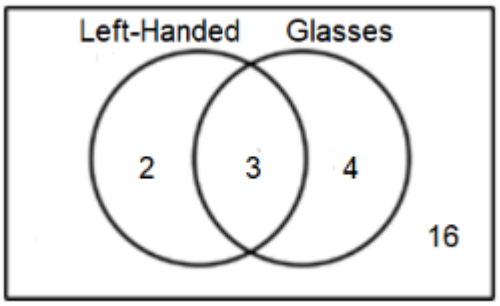
7. a. $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
   b. $\frac{1}{8}$
   c. $\frac{3}{8}$
   d. $\frac{7}{8}$
   e. $\frac{4}{8} = \frac{1}{2}$
8. a. $\{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}$
   b. $\frac{4}{8} = \frac{1}{2}$
   c. $\frac{3}{8}$
   d. $\frac{4}{8} = \frac{1}{2}$
   e. $\frac{2}{8} = \frac{1}{4}$
   f. $\frac{2}{8} = \frac{1}{4}$

Section 5.2

1. $\frac{160}{400} = \frac{2}{5} = 0.4 = 40\%$
2. $\frac{7}{25} = 0.28 = 28\%$
3. a. $\frac{1}{4} = 0.25 = 25\%$
   b. $\frac{6}{40} = 0.15 = 15\%$
   c. red
   d. Probably fair. With only 40 trials, all empirical probabilities are within 10% points of the theoretical probabilities.
4. a. theoretical
   b. experimental
   c. theoretical
   d. experimental
   e. theoretical
   f. experimental
   g. theoretical
   h. experimental

Section 5.3

1. a. $\frac{3}{6} = \frac{1}{2}$
   b. $\frac{2}{6} = \frac{1}{3}$
   c. $\frac{1}{6}$
2. a. 15
   b. 118
   c. $\frac{22}{140} = \frac{11}{70}$



3. a.
   b. $\frac{5}{25} = \frac{1}{5}$
   c. $\frac{18}{25}$
   d. $\frac{4}{25}$
   e. $\frac{9}{25}$
   f. $\frac{16}{25}$
   g. $\frac{3}{5}$
   h. $\frac{2}{18} = \frac{1}{9}$
4. a. $\{a, b\}$

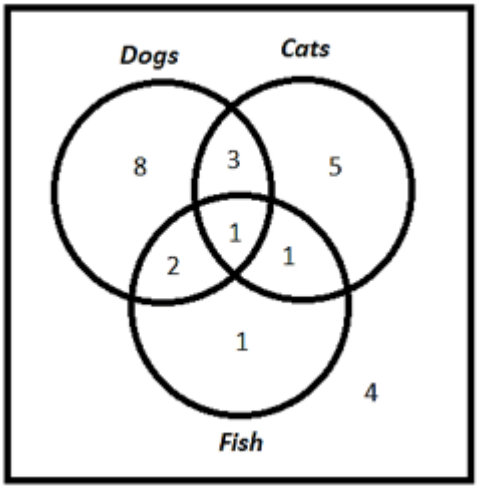
b. $\{c, d\}$
c. $\{a, c, d, e, f\}$
d. $\{\}$ or $\emptyset$
e. $\{f\}$
f. $\{a, b, c, d, e, f\}$

5. a. $\{1, 2, 3, 5\}$
b. $\{1, 2, 3, 4, 5, 6, 7\}$
c. $\{7, 8, 9, 10\}$
d. $\{2, 4\}$
e. $\{1, 2, 3, 4, 5, 6\}$
f. $\{\ \}$ or $\emptyset$
g. $\{1, 2, 3, 5, 7, 8, 9\}$
h. $\{4, 6, 8, 9, 10\}$
i. $\{2\}$
j. $\{8, 9, 10\}$

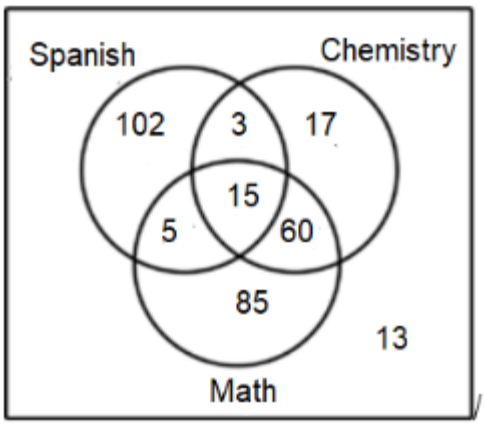
6. a. x, y
b. x, y, z
c. y
d. w, y, z
e. w
f. w, z

7. a. $\frac{16}{105}$
b. $\frac{33}{105} = \frac{11}{35}$
c. $\frac{41}{89}$
d. $\frac{21}{26}$

8. a.



b. 4
c. $\frac{2}{25} = 0.08 = 8\%$
d. $\frac{(8+5+1)}{25} = \frac{14}{25} = 0.56 = 56\%$
e. $\frac{2}{10} = 0.2 = 20\%$



9. a.
   b. i. $\frac{102}{300} = \frac{17}{50}$
      ii. $\frac{60}{300} = \frac{1}{5}$
      iii. $\frac{13}{300}$
      iv. $\frac{204}{300} = \frac{17}{25}$

Section 5.4

1. a. $\frac{22}{110} = \frac{1}{5}$
   b. $\frac{20}{110} = \frac{2}{11}$
   c. $\frac{4}{110} = \frac{2}{55}$
   d. $\frac{38}{110} = \frac{19}{55}$
   e. $\frac{4}{20} = \frac{1}{5}$
   f. $\frac{4}{22} = \frac{2}{11}$
   g. Yes. $P(C) = P(C|T) = \frac{1}{5}$ and $P(T) = P(T|C) = \frac{2}{11}$
2. a. $\frac{120}{646} \approx 0.1858$
   b. $\frac{120}{215} \approx 0.5581$
   c. $\frac{120}{420} \approx 0.2857$
   d. The sample spaces are different.
   e. No. P(6th grade) is $\frac{215}{646} \approx 0.3328$ but P(6th grade | bus) is $\frac{120}{420} \approx 0.2857$.
3. a. $\frac{90}{200} = 0.45$
   b. $\frac{110}{200} = 0.55$
   c. $\frac{162}{200} = 0.81$
   d. $\frac{38}{200} = 0.19$
   e. $\frac{38}{110} \approx 0.3455$
   f. $\frac{38}{90} \approx 0.4222$
   g. No. P(prefers hilly path) = 0.45 and P(prefers hilly path given female) = 0.3455. It is less likely that a cyclist prefers a hilly path if they are female.
4. a. $\frac{84}{200} = 0.42$
   b. $\frac{66}{200} = 0.33$
   c. $\frac{168}{200} = 0.84$
   d. $\frac{66}{200} = 0.44$
   e. $\frac{18}{50} = 0.36$
   f. No. P(below grade level) = 0.42 and P(below grade level given Book A) = 0.44. It is slightly more likely that a student who uses Book A will score below grade level.
5. a. $\frac{47}{350} = 0.1343$
   b. $\frac{32}{350} = 0.0914$
   c. $\frac{209}{350} = 0.5971$
   d. $\frac{17}{119} = 0.1429$
   e. $\frac{13}{134} = 0.0970$
6. 0.4
7. No. P(club) $= 0.40$ and P(club given work) $= \frac{0.05}{0.50} = 0.10$

8. a. $\frac{104}{676} \approx 0.1538$
   b. $\frac{100}{112} \approx 0.8920$
   c. $\frac{100}{104} \approx 0.9615$
   d. $\frac{12}{112} \approx 0.1071$
   e. $\frac{660}{676} \approx 0.9763$

Section 5.5

1. a. $\frac{8}{12} = \frac{2}{3}$
   b. $8 : 4 = 2 : 1$
   c. $4 : 8 = 1 : 2$
2. $2 : 9$
3. $1 : 14$
4. a. $\frac{15}{29}$
   b. $15 : 14$
5. $24 : 1$
6. a. $\frac{6}{10} = \frac{3}{5}$
   b. $6 : 4 \ = \ 3 : 2$
7. a. $4 : 7$
   b. $9 : 2$
   c. $\frac{5}{11}$
8. a. $\frac{7}{16}$
   b. $2 : 11$
   c. $3 : 1$
9. $7 : 3$
10. a. $11 : 89$
    b. $91 : 9$
    c. $\frac{7}{31}$
    d. $\frac{37}{123}$

Chapter 6

Section 6.1

1. b, c, d, e, and h cannot be legitimate probability values because all are outside the range of 0 to 1.
2. $\frac{4}{5}$
3. $P(\text{heart}) + P(\text{not heart}) = 1$

   $\frac{13}{52} + P(\text{not heart}) = 1$

   $P(\text{not heart}) = 1 - \frac{13}{52}$

   $P(\text{not heart}) = \frac{39}{52} = \frac{3}{4}$

4. 31%
5. a and c cannot represent the probabilities because the probabilities do not sum to be 1; d cannot represent the probabilities because all probabilities must be between 0 and 1.
6. 33%
7. 28%
8. 0.85

Section 6.2

1. 86%
2. 65%
3. 0.08

4. 90%
5. 80%
6. 70%
7. 10%
8. 0.56
9. a. $\frac{3}{5}$
   b. $\frac{1}{2}$
   c. $\frac{4}{5}$
   d. $\frac{7}{10}$
10. a. mutually exclusive
    b. not mutually exclusive
    c. not mutually exclusive
    d. not mutually exclusive
11. a. $\frac{26}{52} = \frac{1}{2}$
    b. $\frac{16}{52} = \frac{4}{13}$
    c. $\frac{28}{52} = \frac{7}{13}$
12. a. 86%
    b. 14%
    c. 54%
    d. ≈ 19.4%
    e. ≈ 40.6%
    f. no
    g. no
13. No. because $P(A \cup B) \neq P(A) + P(B)$

Section 6.3

1. 14.31%
2. 58.5%
3. 6%
4. $\frac{2}{87}$
5. $\frac{1}{15}$
6. $\frac{5}{28}$
7. a. $\frac{1}{45}$
   b. $\frac{1}{9}$
   c. $\frac{7}{15}$
   d. $\frac{8}{15}$
8. a. independent
   b. not independent
   c. not independent
   d. independent
9. 0.81%
10. a. 51.2%
    b. 0.8%
    c. 99.2%
11. a. ≈ 31.64%
    b. ≈ 1.17%
    c. ≈ 99.61%

Section 6.4

1. a. $\frac{3}{28}$
   b. $\frac{15}{28}$
   c. $\frac{5}{14}$
   d. $\frac{13}{28}$
   e. $\frac{3}{8}$

2. a. 31.5%
   b. 49%
   c. ≈ 78.57%
3. a. No, the probability of playing football changes when it rains vs. it doesn't rain.
   b. 0.12
   c. 0.32
   d. 0.44
   e. ≈ 0.2727
4. a. 18 %
   b. 32%
   c. 56.25%
5. a. $A = \frac{4}{12}\frac{1}{3}$, $B = \frac{3}{11}$, $C = \frac{8}{11}$, $D = \frac{7}{11}$
   b. $\frac{7}{33}$
   c. $\frac{2}{3}$
6. a. $\frac{20}{27}$
   b. $\frac{4}{9}$
   c. $\frac{7}{9}$
7. a. ≈ 8.33%
   b. ≈ 99.89%

Section 6.5

1. 1.7
2. 2.7
3. Win $0.25 per game on average.
4. $168
5. 1.15
6. a. μ ≈ -$0.81 (expect to lose ≈ $0.81 per spin)
   b. μ ≈ $1.1875 per spin
   c. Part a takes into account that you paid $2.00 to play. Part b is the average of the spin amount, not including the amount you paid to play.
7. earn an average $28 per plan sold
8. $40
9. $125 profit
10. $2

Chapter 7

Section 7.1

1. 18
2. 6
3. 60
4. 1,944
5. 12
6. a. 2,600,000
   b. 786,240
7. a. 2,600,000
   b. 786,240
8. 5,040
9. 40,320
10. 160
11. a. 4
    b. 8
    c. 32
    d. 2^n

Section 7.2

1. a. 72
   b. 210
   c. 220
   d. 720
   e. 3,360
2. 336
3. 10
4. 2,598,96
5. 9,129,120
6. 286
7. 56
8. 24
9. 816
10. 125,970
11. 524,160
12. 6,188
13. 924
14. 3,528
15. 60
16. 9,000
17. a. 720
    b. 60
    c. 50,400
    d. 280
18. 280

Section 7.3

1. $\frac{1}{24} \approx 0.0417$
2. $\frac{1}{32} = 0.03135)$
3. $\frac{2}{5} = 0.4$
4. a. $\frac{1}{190} \approx 0.0053$
   b. $\frac{1}{380} \approx 0.0026$
5. $\frac{1}{252} \approx 0.0040$
6. $\frac{8}{429} \approx 0.0146$
7. $\frac{7,392}{20,995} \approx 0.3521$
8. a. $\frac{1}{42} \approx 0.0238$
   b. $\frac{10}{21} \approx 0.4762$
9. a. $\frac{33}{66,640} \approx 0.0005$
   b. $\frac{2,162}{54,145} \approx 0.0399$
10. a. $\frac{1}{22} \approx 0.0455$
    b. $\frac{5}{11} \approx 0.4545$

This page titled 8.1: Answers to Section Review Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

8.2: Free Online Graphing Tools

Free Online Graphing Tools

The Texas Instruments (TI) calculator, Microsoft *Excel*, and the statistical software package Minitab are excellent tools for analyzing data and producing high-quality graphs. However, you may not always have access to these programs and still need to create graphs. Fortunately, many free online graphing tools can generate clean, professional-looking displays -- some of which produce graphs that even a TI calculator cannot.

As we explore different ways to display data, I encourage you to experiment with free online graph makers. These tools can be especially helpful to you as a prospective teacher when creating polished graphs for tests, worksheets, and homework assignments. Many online tools allow you to save, download, or email your graphs so you can easily insert them into documents. If a tool does not offer these options, you can always take a screen capture and crop the image as needed.

Below are links to several basic graph types discussed in this chapter. This list is not exhaustive, but the tools included are straightforward and easy to use.

Table of Free Online Graphing Resources

Type of Graph and Link	Source
Bar Graph Maker (1)	Rapid Tables
Bar Graph Maker (2)	Social Science Statistics
Box-and-Whisker Plot Maker	Alcula
Circle Graph Maker	Rapid Tables
Frequency Polygon Maker	Social Science Statistics
Frequency Table Distribution Maker	Social Science Statistics
Histogram Maker	Social Science Statistics
Line Graph (Time Series) Maker	Rapid Tables
Linear Regression Calculator	Alcula
Normal Curve Grapher and Calculator	Statistics Kingdom

Type of Graph and Link	Source
Pictograph Maker	SafeSearch
Scatterplot Maker	Alcula
Stem-and-Leaf Plot Maker	Calculator Soup

This page titled 8.2: Free Online Graphing Tools is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

INDEX

No index terms found.

This page titled Index is shared under a Public Domain license and was authored, remixed, and/or curated by Andy Jones.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Andy Jones.

DETAILED LICENSING

Overview

Title: MAT 1160: Data & Chance

Webpages: 52

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 75% (39 pages)
- Undeclared: 7.7% (4 pages)
- CC BY-SA 3.0: 7.7% (4 pages)
- Public Domain: 3.8% (2 pages)
- CC BY-SA 4.0: 3.8% (2 pages)
- Unknown License: 1.9% (1 page)

By Page

- MAT 1160: Data & Chance - *Undeclared*
  - Front Matter - *CC BY-NC-SA 4.0*
    - Title Page - *CC BY-NC-SA 4.0*
    - InfoPage - *Public Domain*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Preface - *CC BY-NC-SA 4.0*
  - 1: An Introduction to Statistics - *CC BY-NC-SA 4.0*
    - 1.1: What Is Statistics? - *CC BY-NC-SA 4.0*
    - 1.2: Populations and Samples - *CC BY-NC-SA 4.0*
    - 1.3: Levels of Measurement - *CC BY-NC-SA 4.0*
    - 1.4: Collecting Data - *CC BY-NC-SA 4.0*
  - 2: Summarizing Data Numerically - *CC BY-NC-SA 4.0*
    - 2.1: Measures of Center - *CC BY-NC-SA 4.0*
    - 2.2: Measures of Position - *CC BY-NC-SA 4.0*
    - 2.3: Measures of Spread - *CC BY-NC-SA 4.0*
    - 2.4: Changes to a Data Set - *CC BY-NC-SA 4.0*
  - 3: Visualizations of Data - *CC BY-NC-SA 4.0*
    - 3.1: Graphs of Categorical Data - *CC BY-NC-SA 4.0*
    - 3.2: Histograms, Frequency Polygons, and Ogives - *CC BY-SA 3.0*
    - 3.3: Other Quantitative Graphs and Data Plots - *CC BY-NC-SA 4.0*
    - 3.4: Box-and-Whisker Plots - *CC BY-SA 3.0*
    - 3.5: Normal Distribution and Its Empirical Rule - *CC BY-SA 4.0*
    - 3.6: Applying a Normal Distribution - *CC BY-NC-SA 4.0*
  - 4: Correlation and Regression - *CC BY-NC-SA 4.0*
    - 4.1: Scatterplots and Correlation - *CC BY-NC-SA 4.0*
    - 4.2: Interpreting a Least Squares Regression Model - *CC BY-NC-SA 4.0*
    - 4.3: Building a Least Squares Model - *CC BY-NC-SA 4.0*
  - 5: An Introduction to Probability - *CC BY-NC-SA 4.0*
    - 5.1: Theoretical Probability - *CC BY-NC-SA 4.0*
    - 5.2: Empirical Probability and the Law of Large Numbers - *CC BY-NC-SA 4.0*
    - 5.3: Using Sets and Venn Diagrams to Represent Events - *CC BY-SA 3.0*
    - 5.4: Contingency Tables, Conditional Probability, and Independence - *CC BY-SA 3.0*
    - 5.5: Odds - *CC BY-NC-SA 4.0*
  - 6: Probability Rules and Expectation - *CC BY-NC-SA 4.0*
    - 6.1: Fundamental Properties of Probability - *CC BY-NC-SA 4.0*
    - 6.2: The Additive Rules of Probability - *CC BY-NC-SA 4.0*
    - 6.3: The Multiplicative Rules of Probability - *CC BY-NC-SA 4.0*
    - 6.4: Tree Diagrams and Multistage Probability - *CC BY-NC-SA 4.0*
    - 6.5: Expectation Involving Chance - *Unknown License*
  - 7: Counting Techniques - *CC BY-NC-SA 4.0*
    - 7.1: The Fundamental Counting Principle - *CC BY-SA 4.0*
    - 7.2: Permutations and Combinations - *CC BY-NC-SA 4.0*
    - 7.3: Probability Using Counting Techniques - *CC BY-NC-SA 4.0*
  - 8: Appendix- Answers and Free Graphing Tools - *CC BY-NC-SA 4.0*

        - 8.1: Answers to Section Review Exercises - *CC BY-NC-SA 4.0*
        - 8.2: Free Online Graphing Tools - *CC BY-NC-SA 4.0*
    - Back Matter - *CC BY-NC-SA 4.0*
        - Index - *Public Domain*
        - Glossary - *CC BY-NC-SA 4.0*
        - Detailed Licensing - *CC BY-NC-SA 4.0*
        - Detailed Licensing - *Undeclared*

This page titled MAT 1160: Data & Chance is shared under a license and was authored, remixed, and/or curated by Andy D. Jones.

DETAILED LICENSING

Overview

Title: MAT 1160: Data & Chance

Webpages: 52

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 75% (39 pages)
- Undeclared: 7.7% (4 pages)
- CC BY-SA 3.0: 7.7% (4 pages)
- Public Domain: 3.8% (2 pages)
- CC BY-SA 4.0: 3.8% (2 pages)
- Unknown License: 1.9% (1 page)

By Page

- MAT 1160: Data & Chance - *Undeclared*
  - Front Matter - *CC BY-NC-SA 4.0*
    - Title Page - *CC BY-NC-SA 4.0*
    - InfoPage - *Public Domain*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Preface - *CC BY-NC-SA 4.0*
  - 1: An Introduction to Statistics - *CC BY-NC-SA 4.0*
    - 1.1: What Is Statistics? - *CC BY-NC-SA 4.0*
    - 1.2: Populations and Samples - *CC BY-NC-SA 4.0*
    - 1.3: Levels of Measurement - *CC BY-NC-SA 4.0*
    - 1.4: Collecting Data - *CC BY-NC-SA 4.0*
  - 2: Summarizing Data Numerically - *CC BY-NC-SA 4.0*
    - 2.1: Measures of Center - *CC BY-NC-SA 4.0*
    - 2.2: Measures of Position - *CC BY-NC-SA 4.0*
    - 2.3: Measures of Spread - *CC BY-NC-SA 4.0*
    - 2.4: Changes to a Data Set - *CC BY-NC-SA 4.0*
  - 3: Visualizations of Data - *CC BY-NC-SA 4.0*
    - 3.1: Graphs of Categorical Data - *CC BY-NC-SA 4.0*
    - 3.2: Histograms, Frequency Polygons, and Ogives - *CC BY-SA 3.0*
    - 3.3: Other Quantitative Graphs and Data Plots - *CC BY-NC-SA 4.0*
    - 3.4: Box-and-Whisker Plots - *CC BY-SA 3.0*
    - 3.5: Normal Distribution and Its Empirical Rule - *CC BY-SA 4.0*
    - 3.6: Applying a Normal Distribution - *CC BY-NC-SA 4.0*
  - 4: Correlation and Regression - *CC BY-NC-SA 4.0*
    - 4.1: Scatterplots and Correlation - *CC BY-NC-SA 4.0*
    - 4.2: Interpreting a Least Squares Regression Model - *CC BY-NC-SA 4.0*
    - 4.3: Building a Least Squares Model - *CC BY-NC-SA 4.0*
  - 5: An Introduction to Probability - *CC BY-NC-SA 4.0*
    - 5.1: Theoretical Probability - *CC BY-NC-SA 4.0*
    - 5.2: Empirical Probability and the Law of Large Numbers - *CC BY-NC-SA 4.0*
    - 5.3: Using Sets and Venn Diagrams to Represent Events - *CC BY-SA 3.0*
    - 5.4: Contingency Tables, Conditional Probability, and Independence - *CC BY-SA 3.0*
    - 5.5: Odds - *CC BY-NC-SA 4.0*
  - 6: Probability Rules and Expectation - *CC BY-NC-SA 4.0*
    - 6.1: Fundamental Properties of Probability - *CC BY-NC-SA 4.0*
    - 6.2: The Additive Rules of Probability - *CC BY-NC-SA 4.0*
    - 6.3: The Multiplicative Rules of Probability - *CC BY-NC-SA 4.0*
    - 6.4: Tree Diagrams and Multistage Probability - *CC BY-NC-SA 4.0*
    - 6.5: Expectation Involving Chance - *Unknown License*
  - 7: Counting Techniques - *CC BY-NC-SA 4.0*
    - 7.1: The Fundamental Counting Principle - *CC BY-SA 4.0*
    - 7.2: Permutations and Combinations - *CC BY-NC-SA 4.0*
    - 7.3: Probability Using Counting Techniques - *CC BY-NC-SA 4.0*
  - 8: Appendix- Answers and Free Graphing Tools - *CC BY-NC-SA 4.0*

        - 8.1: Answers to Section Review Exercises - *CC BY-NC-SA 4.0*
        - 8.2: Free Online Graphing Tools - *CC BY-NC-SA 4.0*
    - Back Matter - *CC BY-NC-SA 4.0*
        - Index - *Public Domain*
        - Glossary - *CC BY-NC-SA 4.0*
        - Detailed Licensing - *CC BY-NC-SA 4.0*
        - Detailed Licensing - *Undeclared*

This page titled MAT 1160: Data & Chance is shared under a license and was authored, remixed, and/or curated by Andy D. Jones.