Math in Society

Mathematics for liberal arts majors

Portland Community College

Pilot Edition 0.3

OpenOregon
Educational Resources

Math in Society

Mathematics for liberal arts majors

Edition: Pilot 0.3

Website: http://spot.pcc.edu/~caralee/Math_105.html

September 23, 2019

Portland Community College

This book was made possible by Open Oregon Educational Resources.

Attributions

Project Lead: Cara Lee
Contributing Authors: Jess Brooks
Cara Lee
Sonya Redmond
Cindy Rochester-Gefre

Thanks to Carlos Cantos and Virginia Somes for input and catching typos.

Cover Image: Portland, Oregon Skyline from the Ross Island Bridge, by Visitor7, used under CC BY-SA 3.0 Unported, cropped.

Chapter 1 is a derivative of Math in Society: Logic, by David Lippman and Morgan Chase, and Math in Society: Sets by David Lippman, used under CC-BY-SA 3.0.

Sections 2.2-2.4 are a derivative of Math in Society: Finance, by David Lippman, used under CC-BY-SA 3.0.

Sections 2.1 and 2.5 are original to Portland Community College.

Section 2.5, Figure 1: Tax Buckets by John Chesbrough, used under CC-BY-ND-NC 4.0.

Chapter 3 is a derivative of Math in Society: Describing Data and Statistics, by David Lippman, Jeff Eldridge and www.onlinestatbook.com, and www.onlinestatbook.com, by David M. Lane, et al, used under CC-BY-SA 3.0.

Chapter 4 is a derivative of Math in Society: Probability, by David Lippman, used under CC-BY-SA 3.0.

Technology Screenshots:
All spreadsheet screenshots use Microsoft Excel under fair use. If you plan to redistribute this book, please consider whether your use is also fair use.

GeoGebra screenshots are used for non-commercial use under https://www.geogebra.org/license#NonCommercialLicenseAgreement

We dedicate this book to our students
May you have greater ease in paying for college
and grow your proficiency and confidence in math.

Word, PDF and Print Versions

This book is available free online at http://spot.pcc.edu/~caralee/Math_105.html. There are Microsoft Word documents and PDF versions of each chapter. There is a PDF version of the required chapters (1-4) online and available at the bookstore for the cost of printing. Only the required chapters are in the printed version, to make the print version cheaper.

This course includes one or more instructor choice topics which may be accessed either from the website above, or http://www.opentextbookstore.com/mathinsociety/index.html.

Accessibility

The word version of each chapter is accessible for use with screen readers. The accessible features include heading navigation, MathType and alternate text on all images. For truth tables and Venn diagrams, files with detailed figure descriptions and graphics optimized for tactile production can be found on the textbook website listed above. If you find anything that can improve the accessibility of this book, please email cara.lee@pcc.edu.

MyOpenMath

Online homework problems are available for free at https://www.myopenmath.com/.

Philosophy

We emphasize technology, conceptual understanding and communication over rote calculation. However, some manual calculation is important to understand what the technology is doing. We emphasize readily available spreadsheets and GeoGebra throughout the text.

Acknowledgements

We would like to thank Amy Hofer of OpenOregon and the PCC OER steering committee. Thanks also to Kaela Parks and Michael Cantino of Disability Services for their expertise on accessibility and for producing the tactile model files.

Table of Contents

Chapter 1: Logic and Sets

Student Outcomes for this Chapter

Section 1.1: The Language and Rules of Logic

Students will be able to:

- Identify propositions
- Compose and interpret the negation of a statement
- Use logical connectors (and/or) and conditional statements (if, then)
- Use truth tables to find truth values of basic and complex statements

Section 1.2: Sets and Venn Diagrams

Students will be able to:

- Use set notation and understand the null set
- Determine the universal set for a given context
- Use Venn diagrams and set notation to illustrate the intersection, union and complements of sets
- Illustrate disjoint sets, subsets and overlapping sets with diagrams
- Use Venn diagrams and problem-solving strategies to solve logic problems

Section 1.3: Describing and Critiquing Arguments.

Students will be able to:

- Understand the structure of logical arguments by identifying the premise(s) and conclusion.
- Distinguish between inductive and deductive arguments
- Make a set diagram to evaluate deductive arguments
- Determine whether a deductive argument is valid and/or sound

Section 1.4: Logical Fallacies.

Students will be able to:

- Identify common logical fallacies and their use in arguments

Section 1.1 The Language and Rules of Logic

Logic

Logic is the study of reasoning. Our goal in this chapter is to examine arguments to determine their validity and soundness. In this section we will look at propositions and logical connectors that are the building blocks of arguments. We will also use truth tables to help us examine complex statements.

Propositions

A **proposition** is a complete sentence that is either true or false. Opinions can be propositions, but questions or phrases cannot.

> Example 1: Which of the following are propositions?
>
> a. I am reading a math book.
> b. Math is fun!
> c. Do you like turtles?
> d. My cat
>
> The first and second items are propositions. The third one is a question and the fourth is a phrase, so they are not. We are not concerned right now about whether a statement is true or false. We will come back to that later when we examine full arguments.

Arguments are made of one or more propositions (called **premises**), along with a **conclusion**. Propositions may be **negated**, or combined with **connectors** like "and", and "or". Let's take a closer look at how these negations and logical connectors are used to create more complex statements.

Negation (not)

One way to change a proposition is to use its **negation**, or opposite meaning. We often use the word "not" to negate a statement.

> Example 2: Write the negation of the following propositions.
>
> a. I am reading a math book. — Negation: I am not reading a math book.
> b. Math is fun! — Negation: Math is not fun!
> c. The sky is not green. — Negation: The sky is green (or not not green).
> d. Cars have wheels — Negation: Cars do not have wheels.

Multiple Negations

It is possible to use more than one negation in a statement. If you've ever said something like, "I can't not go," you are really saying you must go. It's a lot like multiplying two negative numbers which gives a positive result.

In the media and in ballot measures we often see multiple negations and it can be confusing to figure out what a statement means.

Example 3: Read the statement to determine the outcome of a yes vote.

"Vote for this measure to repeal the ban on plastic bags."

If you said that a yes vote would enable plastic bag usage, you are correct. The ban stopped plastic bag usage, so to repeal the ban would allow it again. This measure has a double negation and is also not very good for the environment.

Example 4: Read the statement to determine the outcome on mandatory minimum sentencing.

"The bill that overturned the ban on mandatory minimum sentencing was vetoed."

In this case mandatory minimum sentencing would not be allowed. The ban would stop it, and the bill to overturn it was vetoed. This is an example of a triple negation.

Logical Connectors (and, or)

When we use the word "and" between two propositions, it connects them to create a new statement that is also a proposition. For example, if you said, "when you go to the store, please get eggs ***and*** cereal," you would be expecting both items. For an *and* statement to be true, the connected propositions must both be true. If even one proposition is false (for instance, you get eggs but not cereal), the entire connected *and* statement is false.

The word "or" between two propositions similarly connects the propositions to create a new statement. In this case, if you said, "please get eggs ***or*** cereal," you would be expecting one or the other (but probably not both). For an *or* statement to be true, at least one of the propositions must be true (or both could be true).

Exclusive vs. Inclusive *or*

In English we often mean for *or* to be **exclusive**: one or the other, but not both. In math, however, *or* is usually **inclusive**: one or the other, or both. The thing we are including, or excluding is the "both" option.

Example 5: Determine whether each *or* statement is inclusive or exclusive.

a. Would you like a chicken or vegan meal?
b. We want to hire someone who speaks Spanish or Chinese.
c. Are you going to wear sandals or tennis shoes?
d. Are you going to visit Thailand or Vietnam on your trip?

The first *or* statement is a choice of one or the other, but not both, so it is exclusive. The second statement is inclusive because they could find a candidate

who speaks both languages. The third statement is exclusive because you can't wear both at the same time. The fourth statement is inclusive because you could visit both countries on your trip.

Conditional Statements (if, then)

A **conditional statement** connects two propositions with *if, then*. An example of a conditional statement would be "*If* it is raining, *then* we'll go to the mall."

The statement "If it is raining," may be true or false for any given day. If the condition is true, then we will follow the course of action and go to the mall. If the condition is false, though, we haven't said anything about what we will or won't do.

Basic Truth Tables

In logic we can use a **truth table** to analyze a complex statement by summarizing all the possibilities and their **truth values** (true or false). To do this, we break the statement down to its smallest elements, the propositions. Then we can see the outcome of the complex statement for all possible combinations of true and false for the propositions.

For example, let's work with two propositions:

R: You paid your rent this month.
E: You paid your electric bill this month.

We will use these two propositions to demonstrate the truth tables for *not*, *and*, and *or*.

To set up a truth table, we list all the possible truth value combinations in a systematic way. The standard way of doing this is to make the first column half true, then half false, then cut the pattern in half with each succeeding column. For two propositions, the first two columns are shown to the right.

Truth Table Setup for Two Propositions

R	*E*	
T	T	
T	F	
F	T	
F	F	

The four possible combinations are

Row 1: You have paid your rent and electric bill
Row 2: You have paid your rent but not your electric bill
Row 3: You have not paid your rent but you have paid your electric bill
Row 4: You haven't paid either your rent or electric bill (yet).

Once we fill in the starting columns, we add additional columns for the more complex statements. We can add as many columns as needed. Below are the basic truth tables for *not*, *and,* and *or*.

Basic Truth Tables

Not

R	*not R*
T	F
F	T

In the *not R* column, the truth value is the opposite of the value for *R*. For example, if *R* is true (you paid your rent) then *not R* (you did not pay your rent) is false.

And

R	*E*	*R and E*
T	T	T
T	F	F
F	T	F
F	F	F

In the *R and E* column, you must have paid both your rent and electric bill. Otherwise R *and E* is false.

Or

R	*E*	*R or E*
T	T	T
T	F	T
F	T	T
F	F	F

In the *R or E* column, you must have paid either your rent or electric bill, or both (inclusive or). Otherwise *R or E* is false.

Conditional Truth Tables

We talked about conditional statements (*if, then* statements), earlier. In logical arguments the first part (the *"if"* part) is usually a **hypothesis** and the second part (the *"then"* part) is a **conclusion**.

To understand the truth table values for a conditional statement it is helpful to look at an example. Let's say a friend tells you, "If you post that photo to Facebook, you'll lose your job." Under what conditions can you say that your friend was wrong?

There are four possible outcomes:

1. You post the photo and lose your job
2. You post the photo and don't lose your job
3. You don't post the photo and lose your job
4. You don't post the photo and don't lose your job

The only case where you can say your friend was wrong is the second case, in which you post the photo but still keep your job.

Your friend didn't say anything about what would happen if you didn't post the photo, so you can't say the last two statements are wrong. Even if you didn't post the photo and lost your job anyway, your friend never said that you were guaranteed to keep your job if you didn't post it.

The four cases above correspond to the four rows of the truth table. For this truth table we will use P for "posting the photo," and L for "losing your job."

Truth table for a conditional statement

P	*L*	If *P*, then *L*
T	T	T
T	F	F
F	T	T
F	F	T

If the hypothesis (the "if" part) is false, we cannot say that the statement is a lie, so the result of the third and fourth rows is true. Notice that we are using a double negation in this explanation.

We are using the words *and*, *or*, *not* and *if then* in this book, but if you look up other resources on truth tables you are likely to see these symbols.

Symbols used in other resources

A and B is written $A \wedge B$ — *A or B* is written $A \vee B$

not A is written $\sim A$ — *If A, then B is written* $A \rightarrow B$

Truth Tables for Complex Statements

Truth tables really become useful when we analyze more complex statements. In this case we will have several columns. It helps to work from the inside out and create a column in the table for each intermediate statement.

Example 6: Create a truth table for the statement *A or not B*

When we create the truth table, we start with columns for the propositions, *A* and *B*. Then we add a column for *not B* because that is part of the final statement. Our last column is the final statement *A or not B*.

A	*B*	*not B*	*A or not B*
T	T		
T	F		
F	T		
F	F		

To complete the third column, *not B*, we take the opposite of the *B* column. Then to complete the fourth column, we only look at the *A* and the *not B* columns and compare them using *or*.

A	*B*	*not B*	*A or not B*
T	T	F	T
T	F	T	T
F	T	F	F
F	F	T	T

Truth Tables with Three Propositions

To create a truth table with three propositions we need eight rows for all the possible combinations. We will first determine the columns we need to get to our final statement. Then we will fill in the first three columns using the same methodology as before. Start with half true, half false, then cut the pattern in half each time.

Example 7: Create a truth table for the statement *A and not (B or C)*

First let's figure out the columns we will need. We have *A*, *B*, *C*, then we need the statement in the parentheses, *(B or C)*. Then we need the negation of that column, *not (B or C)*. Then we conclude with our final statement, *A and not (B or C)*.

Here is the initial table:

A	*B*	*C*	*B or C*	*not (B or C)*	*A and not (B or C)*
T	T	T			
T	T	F			
T	F	T			
T	F	F			
F	T	T			
F	T	F			
F	F	T			
F	F	F			

Now we complete the columns one at a time. We use the *B* column and *C* column to complete *B or C*. Then *not (B or C)* is the opposite of that column. For the final column we only need to look at the first and fifth columns, shaded in blue, with *and*. Here is the completed table.

A	*B*	*C*	*B or C*	*not (B or C)*	*A and not (B or C)*
T	T	T	T	F	F
T	T	F	T	F	F
T	F	T	T	F	F
T	F	F	F	T	T
F	T	T	T	F	F
F	T	F	T	F	F
F	F	T	T	F	F
F	F	F	F	T	F

For this statement A must be true and neither B or C can be true, so it is only true in the fourth row. For an example of this statement, let's define these propositions in the context of professional baseball:

Let A = Anaheim wins, B = Baltimore wins, C = Cleveland wins.

Suppose that Anaheim will make the playoffs if: (1) Anaheim wins, and (2) neither Boston nor Cleveland wins. TFF is the only scenario in which Anaheim will make the playoffs.

Example 8: Construct a truth table for the statement *if m and not p, then r.*

First, it may help to add parentheses to help you clarify the order. Our statement could also be written, *if (m and not p), then r*. To build this table, we will build the statement in parentheses and then repeat the *r* column after it. It's easier to read the conditional statement from left to right. Here are the columns for the table:

m	*p*	*r*	*not p*	*m and not p*	*r*	*If (m and not p), then r*
T	T	T				
T	T	F				
T	F	T				
T	F	F				
F	T	T				
F	T	F				
F	F	T				
F	F	F				

For the fourth column, we take the opposite of *p*. Then we use the first and fourth columns to complete *m and not p*. With the *r* column repeated we can use columns five and six to complete our conditional statement. Here is the completed table:

m	*p*	*r*	*not p*	*m and not p*	*r*	*If (m and not p), then r*
T	T	T	F	F	T	T
T	T	F	F	F	F	T
T	F	T	T	T	T	T
T	F	F	T	T	F	F
F	T	T	F	F	T	T
F	T	F	F	F	F	T
F	F	T	T	F	T	T
F	F	F	T	F	F	T

When m is true, p is false, and r is false—the fourth row of the table—then the hypothesis m and not p will be true, but the conclusion is false, resulting in an invalid conditional statement; every other case gives a true result.

> If you want a real-life situation that could be modeled by if m and not p, then r, consider this:
>
> Let m = we order meatballs, p = we order pasta, and r = Ruba is happy.
>
> The statement if *m and not p*, then r is, "if we order meatballs and don't order pasta, then Ruba is happy". If m is true (we order meatballs), p is false (we don't order pasta), and r is false (Ruba is not happy), then the statement is false, because we satisfied the premise, but Ruba did not satisfy the conclusion.

In this section we have discussed propositions, logical connectors and truth tables. In the next section, we will look at set relationships before we analyze arguments.

Exercises 1.1

1. Which of the following are propositions?
    a. Pigs can fly.
    b. What?
    c. I don't know.
    d. I like tofu.

2. Which of the following are propositions?
    a. How far?
    b. Portland is not in Oregon.
    c. Portland Community College.
    d. It is raining.

3. Write the negation of each proposition.
    a. I ride my bike to campus.
    b. Portland is not in Oregon.

4. Write the negation of each proposition.
    a. You should see this movie.
    b. Lashonda is wearing blue.

5. Write a proposition that contains a double negative.

6. Write a proposition that contains a triple negative.

7. For each situation, decide whether the "or" is most likely exclusive or inclusive.
    a. An entrée at a restaurant includes soup or salad.
    b. You should bring an umbrella or a raincoat with you.

8. For each situation, decide whether the "or" is most likely exclusive or inclusive.
    a. We can keep driving on I-5 or get on I-405 at the next exit.
    b. You should save this document on your computer or a flash drive.

9. Translate each statement from symbolic notation into English sentences. Let *A* represent "Elvis is alive" and let *K* represent "Elvis is the King".
    a. *Not A*
    b. *A or K*
    c. *Not A and K*
    d. *If K, then not A*

10. Translate each statement from symbolic notation into English sentences. Let *A* represent "It rains in Oregon" and let *B* represent "I own an umbrella".
    a. *Not B*
    b. *A and not B*
    c. *If A, then B*
    d. *If not B, then A*

Create a Truth Table for each statement.

11. *A and not B*

12. *Not (not A or B)*

13. *Not (A and B and C)*

14. *Not A or (not B and C)*

15. *Not (A and B) or C*

16. *(A or B) and (A or C)*

17. *If A and B, then C*

18. *If A or B, then not C*

19. *If A and C, then not A*

20. *If B or C, then (A and B)*

Section 1.2 Sets and Venn Diagrams

Sets

It is natural for us to classify items into groups, or **sets**, and consider how they interact with each other. In this section, we will use sets and Venn diagrams to visualize relationships between groups and represent survey data.

A **set** is a collection of items or things. Each item in a set is called a **member** or **element**.

Example 1:

a. The numbers 2 and 42 are elements of the set of all even numbers.

b. MTH 105 is a member of the set of all courses you are taking.

A set consisting entirely of elements of another set is called a **subset.** For instance, the set of numbers 2, 6, and 10 is a subset of the set of all even numbers.

Some sets, like the set of even numbers, can be defined by simply describing their contents. We can also define a set by listing its elements using **set notation**.

Set Notation

Set notation is used to define the contents of a set. Sets are usually named using a capital letter, and its elements are listed *once* inside a set of curly brackets.

For example, to write the set of primary colors using set notation, we could name the set C for colors, and list the names of the primary colors in brackets: $C = \{\text{red, yellow, blue}\}$. In this case, the set C is a subset of all colors. If we wanted to write the list of our favorite foods using set notation, we could write $F = \{\text{cheese, raspberries, wine}\}$. And yes, wine is definitely an element of some food group!

Example 2: Julia, Keenan, Jae and Colin took a test. They got the following scores: 70, 95, 85 and 70. Let P be the set of test takers and S be the set of test scores. List the elements of each set using set notation.

In this example, the set of people taking the test is $P = \{\text{Julia, Keenan, Jae, Colin}\}$, and the set of test scores is $S = \{70, 85, 95\}$. Notice in this example that even though two people scored a 70 on the test, the score of 70 is only listed once.

It is important to note that when we write the elements of a set in set notation, there is no order implied. For example, the set {1, 2, 3} is equivalent to the set {3, 1, 2}. It is conventional, however, to list the elements in order if there is one.

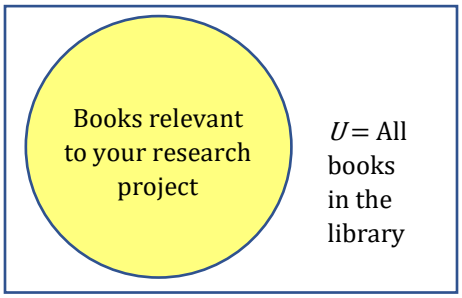
The Universal Set

The **universal set** is the set containing every possible element of the described context. Every set is a therefore a subset of the universal set. The universal set is often

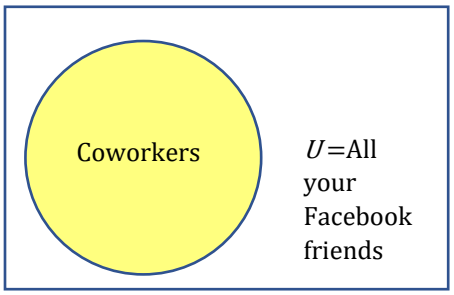
illustrated by a rectangle labeled with a capital letter U. Subsets of the universal set are usually illustrated with circles for simplicity, but other shapes can be used.

Example 3:

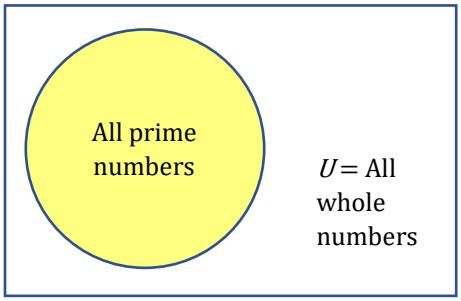
a. If you are searching for books for a research project, the universal set might be all the books in the library, and the books in the library that are relevant to your research project would be a subset of the universal set.



b. If you are wanting to create a group of your Facebook friends that are coworkers, the universal set would be all your Facebook friends and the group of coworkers would be a subset of the universal set.



c. If you are working with sets of numbers, the universal set might be all whole numbers, and all prime numbers would be a subset of the universal set.



The Null Set

It is possible to have a set with nothing in it. This set called the **null set** or **empty set**. It's like going to the grocery store to buy your favorite foods and realizing you left your wallet at home. You walk away with an empty bag. The set of items that you bought at the grocery store would written in set notation as $G = \{\ \}$, or $G = \emptyset$.

Intersection, Union, and Complement (And, Or, Not)

Suppose you and your roommate decide to have a house party, and you each invite your circle, or set, of friends. When you combine your two sets of friends, you discover that you have some friends in common.

The set of friends that you have in common is called the **intersection**. The **intersection** of two sets contains only the elements that are in both sets. To be in the intersection of set *A* and *B*, an element needs to be in both set *A* **and** set *B*.

The set of all friends that you and your roommate have invited is called the **union.** The **union** of two sets contains all the elements contained in either set (or both). To be in the union of set *A* and set *B*, an element must to be contained in just set *A*, just set *B*, **or** in the intersection of sets *A* and *B*. Notice that in this case that the *or* is inclusive.

What about the people who were *not* invited to the party and showed up anyway? They are not elements of your set of invited friends. Nor are they an element of your roommate's set of invited friends. These uninvited party crashers are the **complement** to your set of invited friends. The complement of a set *A* contains everything that is *not* in the set *A*. To be in the complement of set *A,* an element **cannot** be in set *A*, but it will be an element of the universal set.

Example 4: Consider the sets: $A = \{\text{red, green, blue}\}$, $B = \{\text{red, yellow, orange}\}$, and $C = \{\text{red, orange, yellow, green, blue, purple}\}$

a. Determine the set *A* intersect *B*, and write it in set notation.

   The intersection contains the elements in both sets: *A* intersect $B = \{\text{red}\}$

b. Determine the set *A* union *B*, and write it in set notation.

   The union contains all the elements in either set: *A* union $B = \{\text{red, green, blue, yellow, orange}\}$. Notice we only list red once.

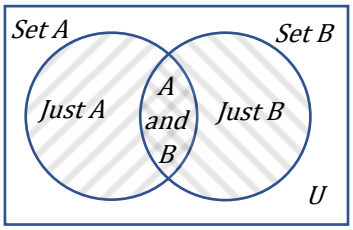
c. Determine the intersection of *A* complement and *C* and write it in set notation.

   Here we are looking for all the elements that are *not* in set *A* and are in set *C*: *A* complement intersect $C = \{\text{orange, yellow, purple}\}$

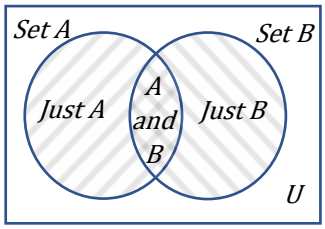
Venn Diagrams

Venn diagrams are used to illustrate the relationships between two or more sets. To create a Venn diagram, start by drawing a rectangle to represent the universal set. Next draw and label overlapping circles to represent each of your sets. Most often there will be two or three sets illustrated in a Venn diagram. Finally, if you are given elements, fill in each region with its corresponding elements.

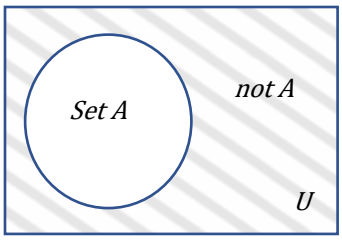
Venn diagrams are also a great way to illustrate intersections, unions and complements of sets as shown below.



The **intersection** is where the shading of the two sets overlaps in the center. It contains the elements of A **and** B.



The **union** includes all elements of A **or** B or both. It contains all three of the shaded regions.



The **complement** of set A includes all the elements **not** in A. It is the shaded region outside the set of A, but within the universal set.

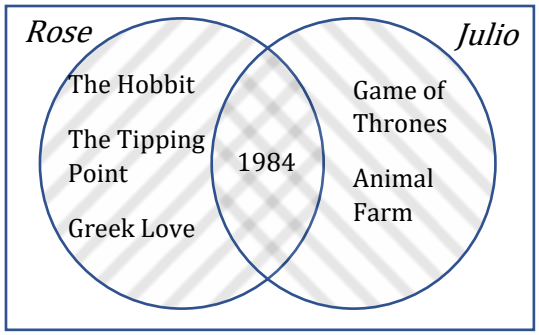
Here is an example of how to draw a Venn Diagram.

Example 5: Let J be the set of books Julio read this summer and let R be the set of books Rose read this summer. Draw a Venn diagram to show the sets of books they read if Julio read Game of Thrones, Animal Farm and 1984, and Rose read The Hobbit, 1984, The Tipping Point, and Geek Love.

To create a Venn diagram showing the relationship between the set of books Julio read and the set of books Rose read, first draw a rectangle to illustrate the universal set of all books.

Next draw two overlapping circles, one for the set of books Julio read and one for the set of books Rose read. Since both Rose and Julio read 1984, we place it in the overlapping region (the intersection).

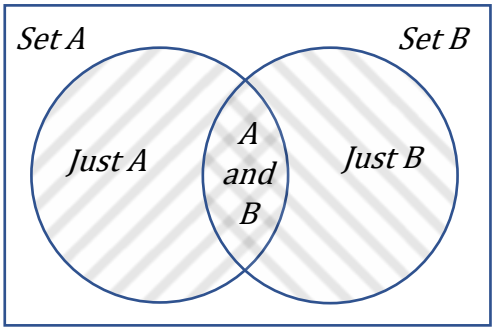
All the books that Rose read will lie in her circle, in one of the two regions that make up her set. Likewise for the books Julio read. Since we have already filled in the overlapping region, we put the books that only Rose read in her circle's "cresent moon" section, and we put the books that only Julio read in his circle's "cresent moon" section. The resulting diagram is shown below.



Example 6: In the last section we discussed the difference between **inclusive "or"** and **exclusive "or**." In common language, "or" is usually exclusive, meaning the set A or B includes just A or just B but not both. In logic, however, "or" is

inclusive, so the set A or B includes just A, just B, or both. The difference between the inclusive and exclusive "or" can be illustrated in a Venn, as shown below.

Inclusive OR includes all three shaded regions, Just A, Just B, and A and B.



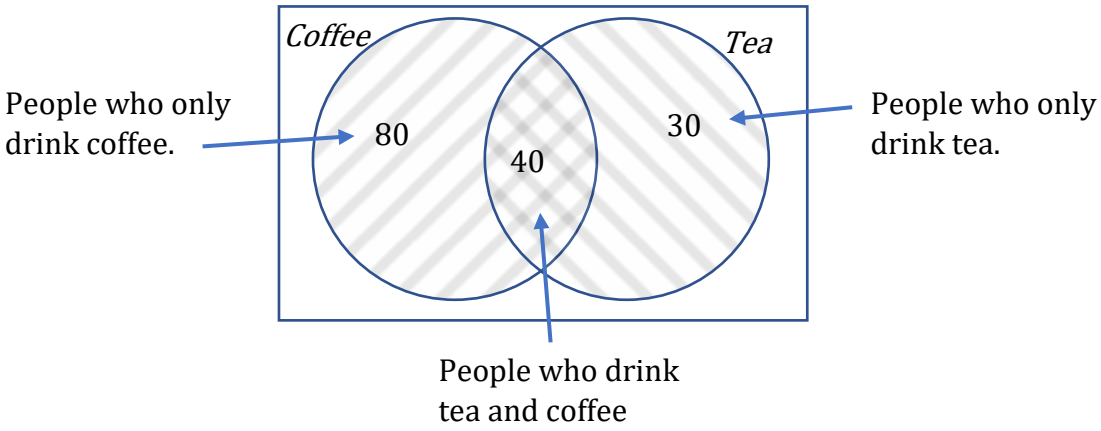
Exclusive OR includes the single shaded regions, Just A and Just B, but not the intersection. You can have A or B but not both.

Illustrating Data

We can also use Venn diagrams to illustrate quantities, data, or frequencies.

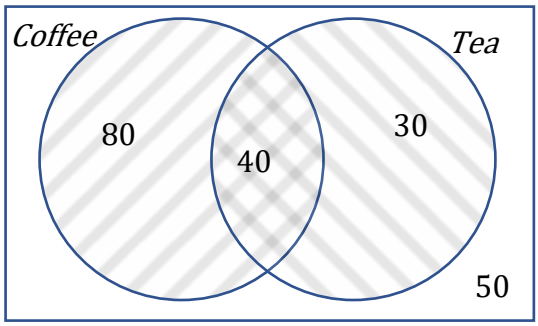
Example 7: A survey asks 200 people, "What beverage(s) do you drink in the morning?" and offers three choices: tea only, coffee only, and both coffee and tea. Thirty report drinking only tea in the morning, 80 report drinking only coffee in the morning, and 40 report drinking both. How many people drink tea in the morning? How many people drink neither tea nor coffee?

To answer this question, let's first create a Venn diagram representing the survey results. Placing the given values, we have the following:



The universal set should include all 200 people surveyed, but we only have 150 placed so far. The difference between what we have placed so far, and the 200 total is the number of people who drink neither coffee nor tea. These 200 – 150 = 50 people are placed outside of the circles but within the rectangle since they are still included in the universal set.

The number of people who drink tea in the morning includes everyone in the tea circle. This includes those who only drink tea and those who drink both tea and coffee. Thus, the number of people who drink tea is 40 + 30 = 70.



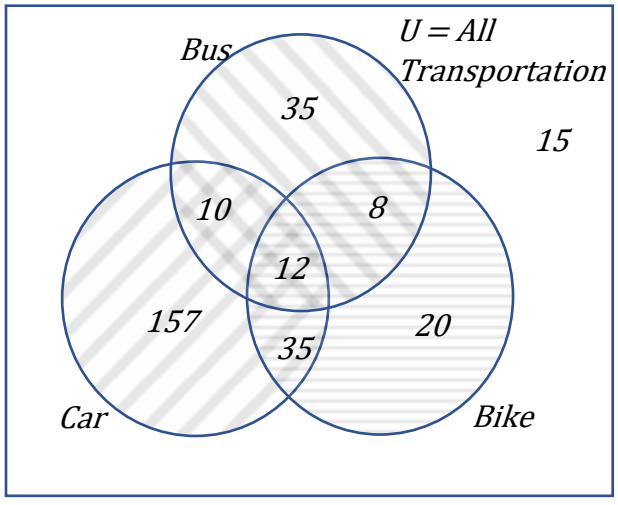
Here is an example of a Venn diagram with three sets.

Example 8: In a survey, adults were asked how they travel to work. Below is the recorded data on how many people took the bus, biked, and/or drove to work. Draw and label a Venn diagram using the information in the table.

Travel Options	Frequency
Just Car	157
Just Bike	20
Just Bus	35
Car and Bike only	35
Car and Bus only	10
Bus and Bike only	8
Car, Bus and Bike	12
Neither Car, Bus nor Bike	15
Total	292

To fill in the Venn diagram, we will place the 157 people who only drive a car in the car set where it does not overlap with any other modes of transportation. We can fill in the numbers 20 and 35 in a similar way.

Then we have the overlap of two modes of transportation only. There are 35 people who use their car and bike only, so they go in the overlap of those two sets, but they do not take the bus, so they are outside of the bus set. Similarly, we can enter the 10 and 8. There are 12 people who use all three modes, so they are in the intersection of all three sets. There are 15 people who do not use any of the three modes, so they are placed outside the circles but inside the universal set of all modes of transportation. Here is the completed Venn diagram.



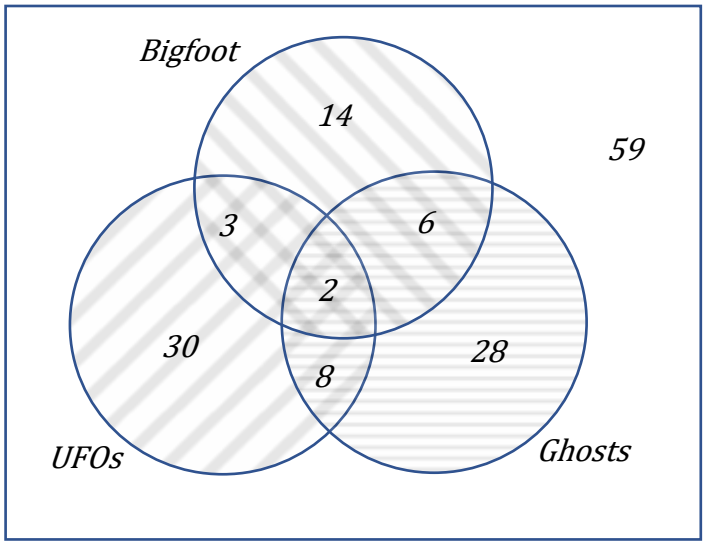
Example 9: One hundred fifty people were surveyed and asked if they believed in UFOs, ghosts, and Bigfoot. The following results were recorded.

- 43 believed in UFOs
- 44 believed in ghosts
- 25 believed in Bigfoot
- 10 believed in UFOs and ghosts
- 8 believed in ghosts and Bigfoot
- 5 believed in UFOs and Bigfoot
- 2 believed in all three

Draw and label a Venn diagram to determine how many people believed in at least two of these things.

Starting with the intersection of all three circles, we work our way out. The number in the center is 2, since two people believe in UFO's, ghosts and Bigfoot. Since 10 people believe in UFOs and Ghosts, and that includes the 2 that believe in all three, that leaves 8 that believe in only UFOs and Ghosts.

We work our way out, filling in all the regions. Once we have, we can add up all those regions, getting 91 people in the union of all three sets. This leaves 150 – 91 = 59 who believe in none.



Then to answer the question of how many people believed in **at least two** (two or more), we add up the numbers in the intersections, 8 + 2 + 3 + 6 = 19 people.

Qualified Propositions

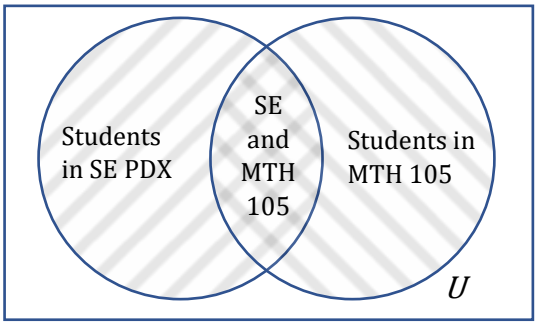
A **qualified proposition** is a statement that asserts a relationship between two sets. The three relationships we will be looking at in this section are "some" (some elements are shared between the two sets), "none" (none of the elements are shared between the two sets), and "all" (all elements of one set are contained in the other set). These relationships are especially important in evaluating arguments.

Overlapping Sets

Sets overlap if they have members in common. The Venn diagram examples we have looked at in this section are **overlapping** sets.

Example 10: The set of students living in SE Portland and the set of students taking MTH 105.

Qualified Proposition: "**Some** students who live in SE Portland take MTH 105."

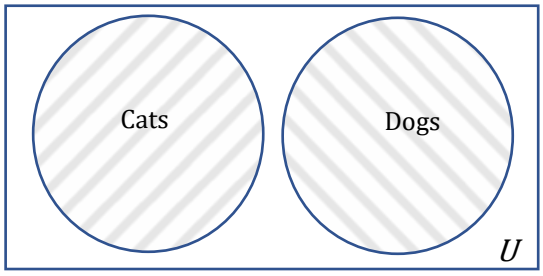


Disjoint Sets

Sets are **disjoint** if they have no members in common.

Example 11: The set of Cats and the set of Dogs.

Qualified Proposition: "**No** cats are Dogs."

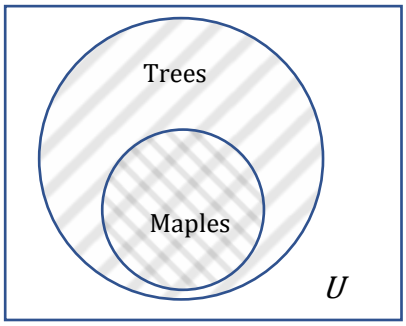


Subsets

If a set is completely contained in another set, it is called a **subset**.

Example 12: The set of all Trees and the set of Maples Trees.

Qualified Proposition: "**All** Maples are Trees."



Exercises 1.2

1. List the elements of the set "The letters of the word Mississippi."

2. List the elements of the set "Months of the year."

3. Write a verbal description of the set {3, 6, 9}.

4. Write a verbal description of the set {a, i, e, o, u}.

5. Is {1, 3, 5} a subset of the set of odd numbers?

6. Is {A, B, C} a subset of the set of letters of the alphabet?

Create a Venn diagram to illustrate each of the following:

7. A survey was given asking whether people watch movies at home from Netflix, Redbox, or a video store. Use the results to determine how many people use Redbox.
   - 52 only use Netflix, 62 only use Redbox
   - 24 only use a video store, 16 use only a video store and Redbox
   - 48 use only Netflix and Redbox, 30 use only a video store and Netflix
   - 10 use all three, 25 use none of these

8. A survey asked buyers whether color, size, or brand influenced their choice of cell phone. The results are below. How many people were influenced by brand?
   - 5 said only color, 8 said only size
   - 16 said only brand, 20 said only color and size
   - 42 said only color and brand, 53 said only size and brand
   - 102 said all three, 20 said none of these

9. Use the given information to complete a Venn diagram, then determine: a) how many students have seen exactly one of these movies, and b) how many have seen only *Star Wars*.
   - 18 have seen *The Matrix* (*M*), 24 have seen *Star Wars (SW)*
   - 20 have seen *Lord of the Rings (LotR),* 10 have seen *M* and *SW*
   - 14 have seen *LotR* and *SW,* 12 have seen *M* and *LotR*
   - 6 have seen all three

10. A survey asked people what alternative transportation modes they use. Use the data to complete a Venn diagram, then determine: a) what percentage of people only ride the bus, and b) what percentage don't use any alternate transportation.
    - 30% use the bus, 20% ride a bicycle
    - 25% walk, 5% use the bus and ride a bicycle
    - 10% ride a bicycle and walk, 12% use the bus and walk
    - 2% use all three

Given the qualified propositions: A) Determine the two sets being described B) Determine if the sets described are Subsets, Overlapping Sets or Disjoint sets. C) illustrate the situation using sets.

11. All Terriers are dogs.

12. Some Mammals Swim. (The second set is not clearly defined but is implied)

13. No pigs can fly.

14. All children are young.

15. Some friends remember your birthday.

16. No lies are truths.

Section 1.3 Describing and Critiquing Arguments

Logical Arguments

A **logical argument** is a claim that a set of **premises** support a **conclusion**. It is possible for a logical argument to have one or many premises, but there must be one conclusion. In this section we will look at types of arguments and how to determine the strength, validity and/or soundness of each type.

There are two types of arguments we will explore in this section: **inductive** and **deductive** arguments.

Inductive and Deductive Arguments

To better understand the difference between inductive and deductive arguments, let's start by looking at a couple of examples.

> Example 1: Consider the following argument:
>
> When I went to the store last week, I forgot my wallet, and I forgot it again when went back today. I always forget my wallet when I go to the store.
>
> Before we analyze an argument, it is helpful to precisely state its premises and its conclusion. Most arguments you encounter in the real world won't be stated in a precise "premise, premise, conclusion" form. Sometimes the conclusion will be stated before the premises, or the premises will be hidden within a bunch of rhetoric.
>
> To begin our analysis of this first argument, let's first rewrite it in a more precise "premise, premise, conclusion" form.
>
> > Premise: I forgot my wallet when I went to the store last week.
> >
> > Premise: I forgot my wallet when I went to the store today.
> >
> > Conclusion: I always forget my wallet when I go to the store.
>
> Notice that both premises make a claim about a **specific** instance – the specific instance last week when I forgot my wallet, and the specific instance today when I forgot my wallet. The conclusion, on the other hand, states what we can expect to happen more generally.

Now let's consider a different argument:

Example 2: Henry must know CPR because he is a nurse and all nurses know CPR.

Just as we did for the last example, let's rewrite the argument in its "premise, premise, conclusion" form:

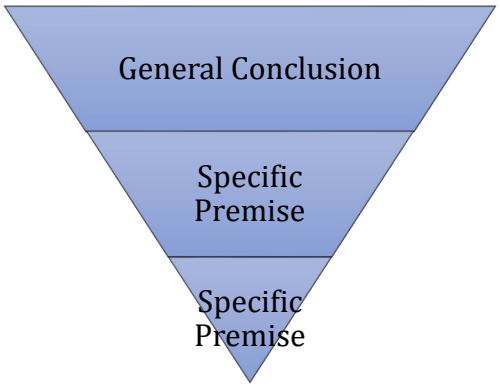
Premise: All nurses know CPR.

Premise: Henry is a nurse.

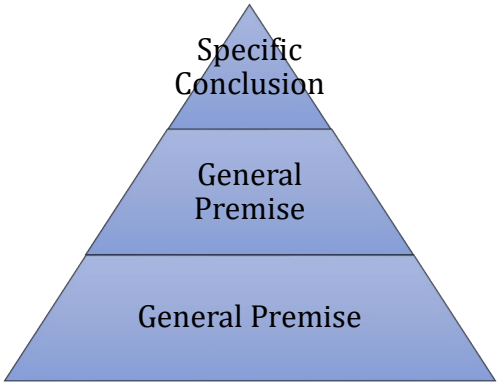
Conclusion: Henry knows CPR.

Unlike the first argument where the premises were specific and the conclusion was general, this argument's first premise is a general statement and the conclusion is specific. We can determine whether an argument is inductive or deductive by looking at which part of the argument is general and which is specific. In the first example, the premises were specific and the conclusion was more general. This is an example of an **inductive** argument. In the second example, it was the premises that were more general and the conclusion that was specific. This is an example of a **deductive** argument.

In general, an **inductive argument** uses a collection of *specific* examples (i.e. data) as its premises and uses them to propose a *general* conclusion, while a **deductive argument** uses a collection of *general* statements (i.e. definitions) as its premises and uses them to propose a *specific* conclusion. You can see the difference in the pyramids below. We start with the premises at the bottom and build up to the conclusion.



Inductive Argument
(Specific to General)



Deductive Argument
(General to Specific)

Example 3:
Rewrite the following arguments in a precise "premise, premise, conclusion" form, and determine if the argument is inductive or deductive.

a. A number is prime if it is only divisible by itself and one. Since the number 13 is only divisible by itself and one, 13 must be prime.

Premise: If a number is only divisible by itself and one, the number is prime.

Premise: The number 13 is only divisible by itself and one.

Conclusion: 13 is prime.

Since the premises are general definitions and properties of numbers and the conclusion is a specific statement about the number 13, the argument is **deductive**.

b. Juan's dog Goober is having puppies. All three of Goober's previous litters have had 5 puppies so Goober is bound to have 5 puppies in this litter as well.

Premise: Goober is having puppies.

Premise: Goober's last three litters had 5 puppies.

Conclusion: Goober's current litter will have 5 puppies.

This is an example of an **inductive argument** since it uses specific experiences/instances as its premises, and its conclusion is a general expectation based on those specific experiences.

Evaluating Arguments

Inductive arguments cannot be proven. The best we can do is evaluate the **strength** of the argument based on the evidence it provides.

A strong inductive argument is one that is well supported by its premises, while a weak inductive argument is one whose premises do a poor job of supporting the conclusion. The strength of an inductive argument is subjective, because where one person sees a strong argument, another may see a weak argument. Additionally, the strength and truth of an argument are not necessarily related; it is possible to have a weak argument that is true, and a strong argument that is false.

Example 4:
Determine the strength of the inductive argument.

James Franco, Jodie Foster, Jennifer Lawrence, and Jack Nicholson have all won Academy Awards for acting. Actors whose names start with J are bound to win an Academy Award.

The inductive argument provides a number of specific cases as evidence for the conclusion. However, we would not be surprised if a J-named actor *did not* win an Academy Award, so the argument is weak

Deductive arguments, on the other hand, can be proven and their validity and soundness can be evaluated. The **validity of the argument** is based on whether the conclusion follows logically from the premises, while the **soundness of the argument** is based on whether or not the premises are true. An argument cannot be sound if it is not valid, even if the premises seem reasonable.

Evaluating Deductive Arguments Using Sets

One way to determine whether a deductive argument is valid is to illustrate the premises of the argument using sets and see if the conclusion logically follows if we assume the premises to be true.

Example 5:
Use a set diagram to determine whether the argument is valid. If the argument is valid, determine if it is also sound.

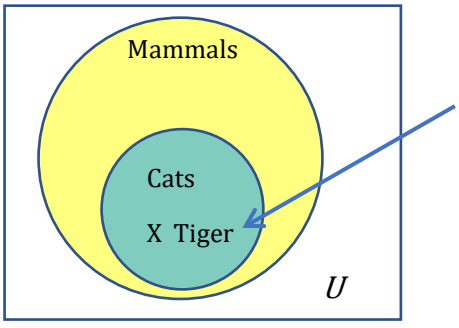
a. "All cats are mammals and a tiger is a cat, so a tiger is a mammal."
First let's write the argument in its "premise, premise, conclusion" form. For the problems we will be looking at, you will want to write the first premise as a **qualified proposition** (some, none, all) since this will form the basic structure of our diagram.

Premise: All cats are mammals.

Premise: A tiger is a cat.

Conclusion: A tiger is a mammal.

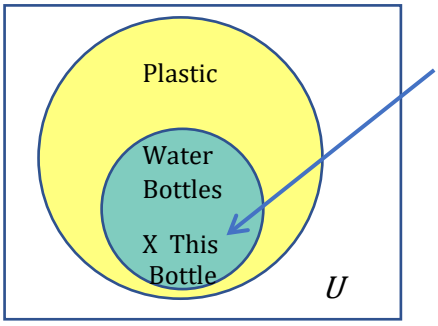
From the first premise we know that all cats lie inside the set of mammals (cats are a subset of mammals). From the second premise, we know that tigers lie inside the set of cats (marked with an X), and therefore also lie within the set of mammals.



Tigers are in the Cats circle, so they must also be in the Mammals circle.

This argument is valid because we were able to show that the conclusion follows logically from the premises. The argument is also sound since the premises "all cats are mammals" and "a tiger is a cat" are true.

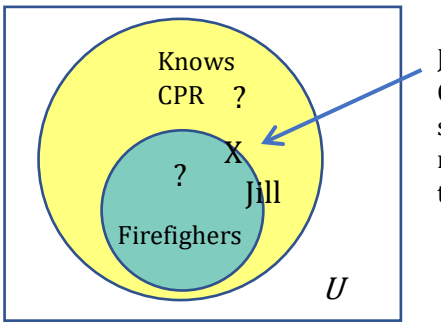
b. "All water bottles are plastic. This is a water bottle, so it must be plastic." From the first premise we know that all water bottles lie inside the set of plastic items (water bottles are a subset of plastic). From the second premise, we know that this particular water bottle must lie within the plastic items set.



Any water bottle will be inside the plastic circle also.

This argument is valid because we were able to show that the conclusion follows logically from the premises. But the argument is not sound because the premise that all water bottles are plastic is not true. There are many versions of glass and metal bottles that are evidence that the first premise is not true. This argument is valid but not sound.

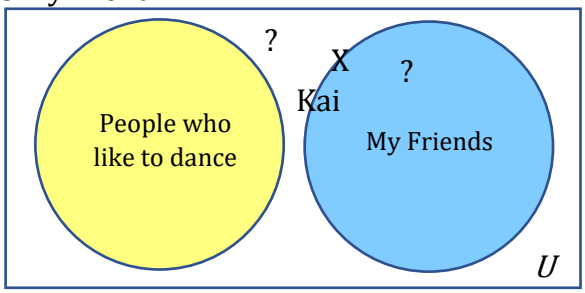
c. "All firefighters know CPR. Jill knows CPR, so Jill must be a firefighter." From the first premise we know that all firefighters lie inside the set of those who know CPR (firefighters are a subset of people who know CPR). From the second premise, we know that Jill is a member of the set of those who know CPR, but we do not have enough information to know whether she is also a member of set of firefighters.



Jill is somewhere in the "Knows CPR" circle, but it is not clear if she is in the firefighter circle or not. We will put the X for Jill on the border of the two regions.

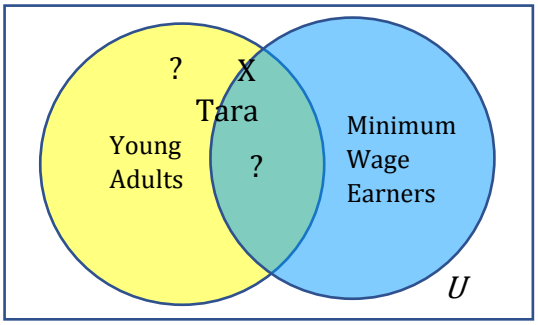
Since we cannot determine which group Jill must be a part of, the argument is invalid. The statement that Jill is a firefighter does not follow logically from the premises that "all firefighters know CPR" and that "Jill knows CPR". Since the argument is not valid, it cannot be sound.

d. "None of my friends like dancing. Kai doesn't like dancing. Therefore, Kai is my friend."



Because it said "none" we draw disjoint sets – one set for my friends and a second set for people who like to dance. The second premise tells us that Kai doesn't like to dance so they're not in the set of people who like to dance. However, we can't put Kai in the set of my friends either. They could be my friend, or someone I don't know who happens to not like dancing. Therefore, the conclusion is **not valid**. And therefore, is also **not sound**.

e. "Some young adults make minimum wage and Tara is a young adult. Therefore, Tara makes minimum wage."



Because it said "some" we draw overlapping sets. The second premise tells us to put Tara in the set of young adults, but it doesn't tell us if she makes minimum wage or not. So, like the previous example we cannot determine which region she is in. She could make minimum wage, or she could also make more. Therefore, the conclusion is **not valid** and therefore, **not sound**.

Exercises 1.3

Rewrite each of the following arguments in their "premise, premise, conclusion" form, and determine whether the argument is inductive or deductive. If the argument is inductive, determine its strength. If the argument is deductive, use sets to illustrate and determine the validity of the argument, and state whether the argument is sound.

1. Since all cats are scared of vacuum cleaners and Max is a cat, Max must be scared of vacuum cleaners.

2. Every day for the last year, a plane flew over my house at 2 pm. Therefore, a plane will always fly over my house at 2pm.

3. Kiran collected data on the salaries of their friends. They found that female and nonbinary friends made less than male friends, so they concluded that women and nonbinary people make less than men.

4. Some of these kids are rude. Jimmy is one of these kids. Therefore, Jimmy is rude!

5. All bicycles have two wheels. My friend's Harley-Davidson has two wheels, so it must be a bicycle.

6. Since all chocolate contains nuts and this bar is made of chocolate, it must contain nuts.

7. All students drink a lot of caffeine. Brayer drinks a lot of caffeine, so he must be a student.

8. Over the course of a year, data was collected on the number of students visiting the Cafeteria. On average, there were 15-35 students present in the cafeteria during the peak hours of the data. We can expect there to be between 15 and 35 students in the cafeteria if we go during the peak hours of the day.

9. If a person is on this reality show, they must be self-absorbed. Laura is not self-absorbed. Therefore, Laura cannot be on this reality show.

For each, draw the appropriate illustration of sets (Subset, Disjoint or Overlapping). Then put an X to represent the subject of the conclusion. Or two question marks to illustrate the subject could into two locations. Finally, state if the conclusion is valid.

10. Premise: No apples are pears.
    Premise: A Pink Lady is an apple.
    Conclusion: Therefore, a Pink Lady is not a pear.

11. Premise: All children are young.
    Premise: Tamika is young.
    Conclusion: Therefore, Tamika is a child.

12. Premise: Some goats faint.
    Premise: Fizzy faints.
    Conclusion: Therefore, Fizzy is a goat.

13. Premise: All students who miss more than 25% of class time fail.
    Premise: Claudia failed my class.
    Conclusion: Claudia missed more than 25% of class time.

14. Premise: All students who miss more than 25% of class time fail.
    Premise: Ethan missed more than 25% of class time.
    Conclusion: Ethan failed.

Section 1.4 Logical Fallacies

Logical Fallacies

In the last section we saw that logical arguments are invalid when the premises are not sufficient to guarantee the conclusion, and that even if an argument is valid it may be unsound if the premises are not true. There are other ways that a logical argument my by invalid or unsound. One of the more common ways this can occur is if the argument is a **fallacy**.

A **fallacy** is a type of argument that appears valid but uses a logical error to persuade or deceive. Fallacious arguments are especially common in advertising and politics, so it is important as informed citizens to recognize when we are being presented with a fallacious argument and to not be persuaded by it.

Common Logical Fallacies

There are many logical fallacies, and some go by more than one name. Below we introduce a few of the more common fallacies that you will be asked to recognize by name, but there are many others.

Personal Attack (Ad hominem)

A **personal attack** argument attacks the person making the argument while ignoring the argument itself. A personal attack is not the same as an insult. Rather, a personal attack claims that there is something wrong with the person or group in order to cast doubt on their character and discredit their argument.

> Example 1: "Jane says that whales aren't fish, but she's only in the second grade so she can't be right."
>
> Here the argument is attacking Jane, not the validity of her claim, so this is a personal attack.
>
> Example 2: "Jane says that whales aren't fish, but everyone knows that they're really mammals. She's so stupid."
>
> This certainly isn't very nice, but it is *not* a personal attack since a valid counterargument is made ("they really are mammals") along with a personal insult.
>
> Example 3: "Mr. Smith is a college dropout, so his stance on education reform cannot be trusted."
>
> Here the argument uses the fact that Mr. Smith did not complete their college degree to discredit their ideas on education reform, so it is a personal attack.

Appeal to Ignorance

An **appeal to ignorance** argument assumes something is true because it hasn't been proven false.

> Example 4: "Nobody has proven that photo isn't of Bigfoot, so it must be Bigfoot."
>
> This is an example of an appeal to ignorance since the fact that no one has been able to prove the picture of Bigfoot is false is being used as evidence that it is Bigfoot.

Appeal to Authority

An **appeal to authority** argument attempts to use the authority of a person to prove a claim. An authority could be an expert such as a doctor or scholar, or someone who is admired like a celebrity or sports figure. While an authority can provide strength to an argument, problems can occur when the person's opinion is not shared by other experts, or when the authority is irrelevant to the claim.

> Example 5: "A diet high in bacon can be healthy; Doctor Atkins said so."
>
> Here, an appeal to a doctor's authority is used for the argument. This generally would provide strength to the argument, except that the opinion that eating a diet high in saturated fat runs counter to general medical opinion. More supporting evidence would be needed to justify this claim.
>
> Example 6: "Jennifer Hudson and Oprah lost weight with Weight Watchers, so their program must work."
>
> In this example there is an appeal to the authority of celebrities. While their experience does provide evidence, it provides no more than any other person's experience would.

False Dilemma

A **false dilemma** argument falsely frames an argument as an "either or" choice without allowing for additional options.

> **Example:** "Either those lights in the sky were an airplane or aliens. There are no airplanes scheduled for tonight, so it must be aliens."
>
> This argument is a false dilemma since it ignores the possibility that the lights could be something other than an airplane or aliens.

Straw Man (or Straw Person)

A **straw person** argument involves misrepresenting the argument in an oversimplified, distorted and less favorable way to make it easier to attack.

> Example 7: "Senator Khouri has proposed reducing military spending by 10%. Apparently, she wants to leave us defenseless against attacks by terrorists."

Here the arguer has represented a 10% funding cut as equivalent to leaving us defenseless, making it easier to attack Senator Khouri's position.

Post Hoc

A **post hoc** argument claims that because two things happened sequentially, then the first must have *caused* the second.

Example 8: "Every morning the rooster crows just before dawn. It must be his crow that makes the sun rise."

Here the arguer is saying the rooster caused the sun to rise, but it is more likely that the sun rising caused the rooster to crow.

Example 9: "Today I wore a red shirt and my football team won! I need to wear a red shirt every time they play to make sure they keep winning."

This person is saying their team won because they wore a red shirt. This type of superstition is quite common in sports even though we really know they are unrelated.

Sometimes there may be more than one fallacy that seems reasonable. Consider this argument: "Emma Watson says she's a feminist, but she posed for these racy pictures. I'm a feminist and no self-respecting feminist would do that." Could this be ad hominem, saying that Emma Watson has no self-respect? Could it be appealing to authority because the person making the argument claims to be a feminist? Could it be a false dilemma because the argument assumes that a woman is either a feminist or not, with no gray area in between?

We have described just six of the many types of logical fallacies. Once you learn to recognize these you will also likely become aware of many others. There are many lists of logical fallacies online.

Exercises 1.4

Determine which type of fallacy each argument represents.

1. John Bardeen's work at the Advanced Institute for Physics has progressed so slowly that even his colleagues call him a plodder. Hence, it is prudent at present not to take seriously his current theory relating how strings constitute the smallest of subatomic particles.

2. You will tell the general manager that I made the right choice in dealing with that customer. After all, I'm the shift manager, so my decisions are always right.

3. It was his fault, Officer. You can tell by the kind of car I'm driving and by my clothes that I am a good citizen and would not lie. Look at the rattletrap he is driving and look at how he is dressed. You can't believe anything a dirty, longhaired hippie like that might tell you. Search his car; he probably has pot in it.

4. We can go to the amusement park or the library. The amusement park is too expensive, so we must go to the library.

5. During the Gulf war many American businesses made immense profits. That is an indisputable fact. Therefore, there can be no doubt that American business interests instigated the war.

6. The oven was working fine until you started using it, so you must have broken it.

7. Old man Brown claims that he saw a flying saucer in his farm, but he never got beyond the fourth grade in school and can hardly read or write. He is completely ignorant of what scientists have written on the subject, so his report cannot possibly be true.

8. There are a number of fallacies that were not discussed in this section. Do an internet search for the following fallacies. Provide both a definition and at least one example.

    a. Slippery Slope
    b. Circular Reasoning
    c. Appeal to Emotion
    d. Red Herring
    e. Whataboutism

Chapter 2: Financial Math

Student Outcomes for this Chapter

Section 2.1: Introduction to Spreadsheets

Students will be able to:

- Perform basic calculations on a spreadsheet
- Use cell references and the fill-down feature

Section 2.2: Simple and Compound Interest

Students will be able to:

- Use spreadsheet functions and/or mathematical formulas to calculate simple, compound, and continuously compounded interest
- Understand the difference between simple and compound interest
- Use a spreadsheet to calculate the effective rate and compare accounts
- Use a spreadsheet and/or formula to calculate the present value needed to reach a desired future value

Section 2.3: Savings Plans

Students will be able to:

- Use a spreadsheet and/or formula to calculate the future value and interest earned on savings plans
- Use a spreadsheet and/or formula to calculate payment amounts for savings plans
- Analyze and compare lump sum and regular payment savings plans

Section 2.4: Loan Payments

Students will be able to:

- Use a spreadsheet and/or formula to calculate the payment amount for student loans, car loans, paying off credit cards and mortgage loans
- Calculate the total paid over the life of a loan, amount of interest paid, and the percentage of the total amount paid in interest
- Determine when to use each formula in the financial math chapter

Section 2.5: Income Taxes

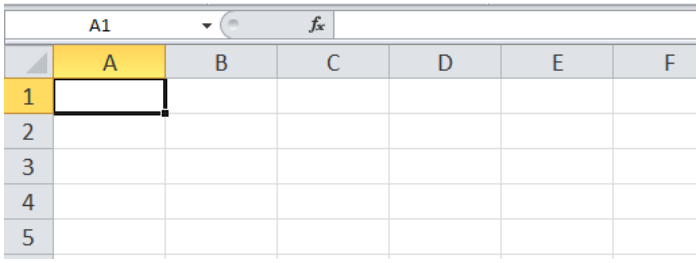
Students will be able to:

- Calculate gross income and adjusted gross income (AGI)
- Determine the standard deduction according to filing status
- Determine whether to use the standard or itemized deductions and calculate taxable income
- Calculate income tax from tables
- Compare taxes owed to withholdings to determine whether a refund is due or a payment is required

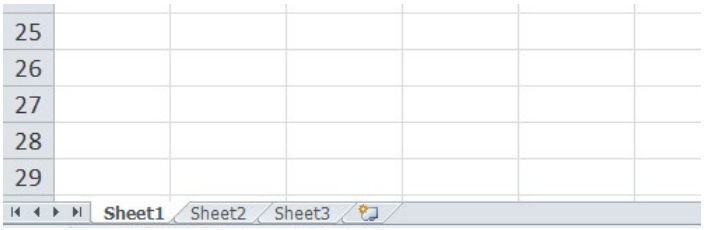
Section 2.1 Introduction to Spreadsheets

A spreadsheet such as Google Sheets or Microsoft Excel, is a very useful tool for doing calculations and making complex tables. You can type in your own custom calculations or use the built-in formulas.

The rectangles within a spreadsheet are called **cells**, and they can be referenced by their column letter and row number. The first cell in the upper left side highlighted below is A1. If we wanted to talk about the third column and the fifth row, that cell would be C5.



A spreadsheet file can contain many sheets. Look along the bottom to see if there is more than one sheet and make sure you are on the right sheet.



Basic Calculations

To do a calculation on a spreadsheet, type an equal sign before the operation. This lets the program know that you want it to calculate the result. When you press enter, you will see the result.

Example 1:

a. To add 3 + 4 =3+4
b. To subtract 100-76 =100-76
c. 4 times 18 =4*18
d. 0.05 divided by 12 =0.05/12
e. To calculate 5^{25} =5^25

	A
1	=3+4
2	=100-76
3	=4*18
4	=0.05/12
5	=5^25

Note that the asterisk (*) is used for multiplication. Spreadsheets don't recognize parentheses as indicators of multiplication like calculators do, so even if you have parentheses for the order of operations, the asterisk is also needed.

You can make more complicated mathematical expressions using parentheses and other operations. To **edit a cell**, click on the editing box at the top, or double click on the cell to edit it directly.

Example 2:

Your bill at a restaurant is $35.75 and you want to leave an 18% tip. How much would you add to the bill?

To work with a percentage, we need to convert it into a decimal first, and then multiply it by the base amount. In a spreadsheet we would type

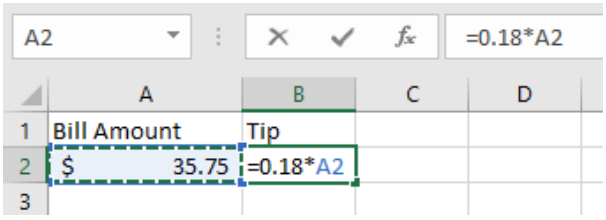
=0.18*35.75

=$6.44, **rounded to the nearest cent**. You would leave a tip of $6.44.

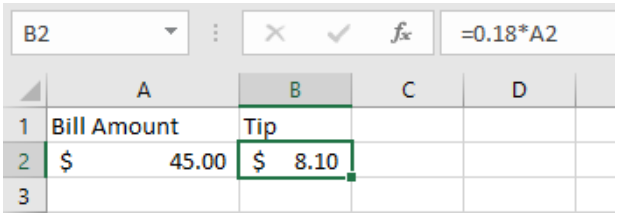
Cell References

One of the powerful things about spreadsheets is using a **cell reference**, such as C5 in a calculation. When you use a cell reference, the values will automatically update if any of the referenced values change.

Let's make a spreadsheet for the percentage tip example above. We calculated an 18% tip on a bill of $35.75. We might want to tip 18% in general, but our bill will change values. We labeled the first column Bill Amount and the second column Tip. The amount of $35.75 is entered in cell A2. Then when we write our formula in B2, we want to calculate 18% of A2. That way if the number in A2 changes, our tip will automatically update.



The formula =0.18*A2 is entered in B2 which gives a result of $6.44 when you hit enter.



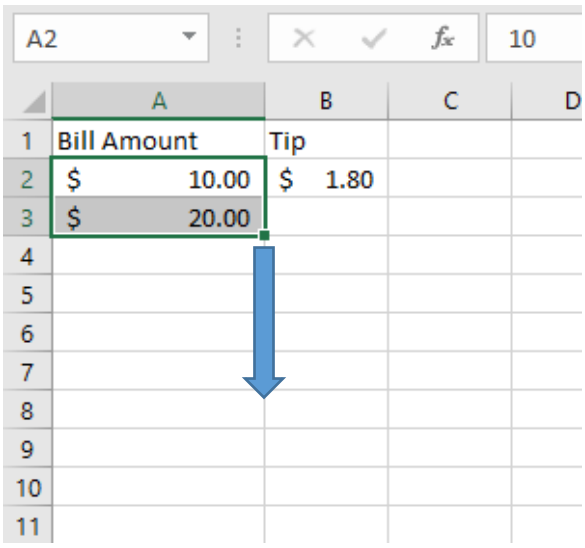
When the bill amount is changed, the tip is recalculated.

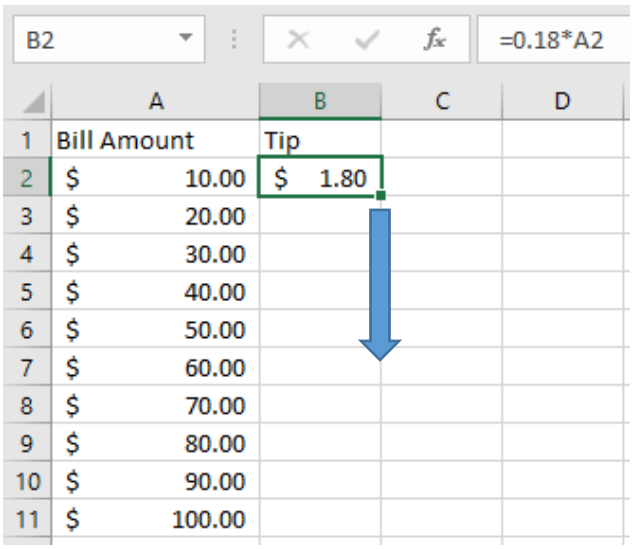
Cell Formatting

We can also format cells A1 and B1 to show dollar signs by clicking on the dollar sign in the number formatting menu.

Fill-Down Feature

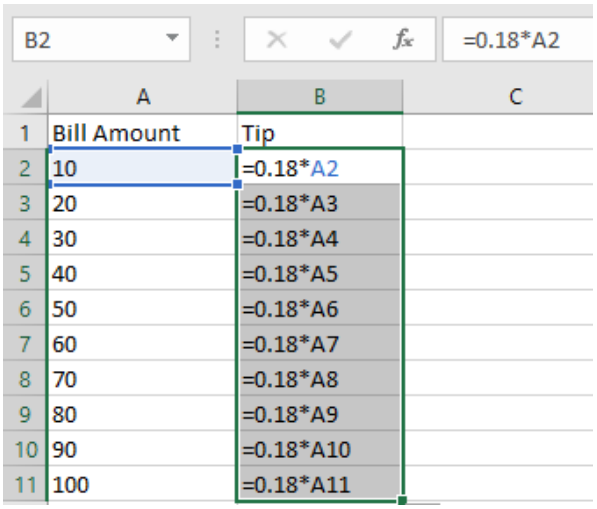
The **fill-down feature** is very useful for making tables. This allows us to copy values or formulas to save time. Let's make a tipping reference table with values from $10, to$100, in increments of $10. First, we will enter two values in column A to establish the pattern. Then select those two cells and you will see a small square in the lower right corner. Drag that square down until you get to $100.



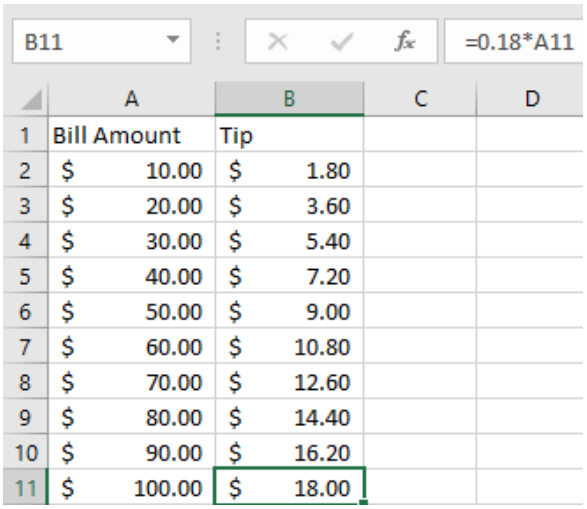


Next, we can drag our formula down and the cell reference will change to each row number automatically.

Here are the formulas with the row numbers updated:



Here is our completed table with the calculations:



Formulas

Spreadsheets have many useful built-in formulas. We will introduce some of the financial formulas in this chapter. Here are some of the formulas we will use:

=FV to calculate the future values of an investment
=PV to calculate the deposit needed for a desired future balance
=PMT to calculate a loan or savings plan payment
=EFFECT to calculate the effective rate of an account and compare accounts

In the rest of this chapter we will use spreadsheets and formulas to calculate the future values, interest paid or earned and monthly payments.

Exercises 2.1

Use a spreadsheet to compute the following.

1. Convert 4/7 to a decimal
2. Convert 16% to a decimal
3. Add 8 and 19
4. Find the difference of 230 and 78
5. Multiple 12 and 9
6. Divide 0.09 by 52
7. Calculate 8^3
8. Your bill at a restaurant is $55.75 and you want to leave a 20% tip. How much would you add to the bill?
9. You leave a tip for $7.50 for a bill at a restaurant that is $44.50. What percent tip did you leave?
10. In Column A use the fill down feature to build a spreadsheet starting with $5 and ending at $125, in increments of $5. In Column B write a formula with a cell reference to calculate a 15.5% tip on the amount in Column A. Use the fill down feature to complete your table.

Section 2.2 Simple and Compound Interest

Note: Spreadsheets are emphasized in this chapter, but the formulas are also presented so you can understand what the spreadsheet is doing. Be sure to check with your instructor for which method to use.

Working with money is a very important skill for everyday life. While balancing a checkbook or calculating our monthly expenditures on espresso requires only arithmetic, when we start saving money, planning for retirement, or need a loan, we need more mathematics and tools. In this section we will calculate and compare simple and compound interest.

Simple Interest

Calculating interest starts with the **principal**, P, or the beginning amount in your account. This is also called the **present value**. This could be a starting investment, or the starting amount of a loan. The **interest**, I, in its most simple form, is a percentage of the principal.

For example, if you borrowed \$100 from a friend and agree to repay it with 5% interest, then the amount of interest you would pay would be 5% of 100. It is very important to remember to change the interest **rate**, r, of 5% into a decimal by moving the decimal two places to the left.

$(0.05)\$100 = \$5$

The total amount you would repay is called the **future value** and would be \$105, the original principal plus the interest.

$\$100 + \$5 = \$105$

Here are the formulas to represent the calculations we just did.

Simple One-time Interest

$$I = Prt$$

$$A = P + I \quad \text{or} \quad A = P + Prt$$

I	is the interest
P	is the principal, starting amount or **present value**
r	is the interest rate in decimal form
t	is the time the money is invested or borrowed for
A	is the end amount: principal plus interest. This is also called the **future value**

Example 1:

A friend asks to borrow \$300 and agrees to repay it in 30 days with 3% simple interest. How much interest will you earn?

$P = \$300$	the principal or present value
$r = 0.03$	3% rate

Using the formula,

$$I = Prt$$
$$= \$300(0.03)$$
$$= \$9$$

To calculate this in a spreadsheet, you would enter

`=300*0.03`

=$9. You will earn $9 in interest when your friend pays you back.

One-time simple interest is only common for extremely short-term or informal loans. For longer term loans or investments, it is common for interest to be paid on a daily, monthly, quarterly, or annual basis. In that case, interest would be earned regularly. Bonds are an example of this type of investment. Bonds are issued by the federal, state or local governments to cover their expenses.

Example 2:

Suppose your city is building a new park, and issues bonds to raise the money to build it. You buy a $1,000 bond that pays 5% simple interest annually and matures in 5 years. How much interest will you earn? What is the future value of the bond?

Each year, you would earn 5% interest so over the course of five years, you would earn:

$$\$1,000(0.05)(5) = \$250$$

When the bond matures, you would receive back the $1,000 you originally paid and the $250 in interest, so we could also put that into a single calculation:

$$\$1,000 + \$1,000(0.05)(5) = \$1,250$$

Using a spreadsheet, you would enter

`=1000+1000*0.05*5`

=$1,250.

The future value of the bond is $1,250.

We can generalize this idea of simple interest over time.

Simple Interest over Time

$$I = Prt$$
$$A = P + I \quad \text{or} \quad A = P + Prt$$

I	is the interest
P	is the principal, starting amount, or present value
r	is the interest rate in decimal form
t	is time, where the increment of time (years, months, etc.) matches the time period for the interest rate
A	is the end amount, principal plus interest, or future value

APR – Annual Percentage Rate

Interest rates are usually stated as an annual percentage rate (APR) – the total interest that will be paid in the year. If the interest is paid in smaller time increments, the APR will be divided by the number of time periods.

For example, a 6% APR paid monthly, would be divided by 12, because you would get one twelfth of the rate per month, which is half a percent per month.

$$\frac{0.06}{12} = 0.005$$

A 4% annual rate paid quarterly, would be divided by 4 to get 1% per quarter.

$$\frac{0.04}{4} = 0.01$$

Here is an example of a semi-annual rate.

Example 3:

Suppose you buy a $1,000 federal bond with a 4% annual simple interest rate, paid semi-annually, with a maturity in 4 years. How much interest will you earn? What will be the future value of the bond?

$P = \$1000$ the principal

$r = \frac{0.04}{2} = 0.02$ interest is being paid semi-annually (twice a year), so the 4% interest will be divided into two 2% payments.

$t = 8$ 4 years, compounded twice a year so $t = 4 \cdot 2 = 8$ half-years

$$I = Prt$$
$$= \$1000(0.02)(8)$$
$$= \$160$$

You will earn $160 in interest over the four years. The future value of the loan is

$$A = P + I$$
$$= \$1000 + \$160$$
$$= \$1{,}160$$

We could also use a spreadsheet to do this calculation and enter:

```
=1000+1000*(0.04/2)*(4*2)
```

=$1,160

The future value of the bond is $1,160. Remember that spreadsheets don't interpret parentheses as multiplication. We need the asterisks as well as the parentheses.

Compound Interest

In a standard bank account, any interest we earn is automatically added to our balance, and we earn interest on that interest. This reinvestment of interest is called

compounding. We will develop the mathematical formula for compound interest and then show the equivalent spreadsheet function.

Suppose that we deposit $1000 in a bank account offering 3% interest, compounded monthly. How will our money grow?

The 3% interest is an annual percentage rate (APR) – the total interest to be paid during the year. Since interest is being paid monthly, each month, we will earn $\frac{0.03}{12} = 0.0025$ per month.

In the first month,

$P = \$1000$
$r = 0.0025$ (which is 0.25%)
$I = \$1000(0.0025) = \$2.50$
$A = \$1000 + \$2.50 = \$1002.50$

In the first month, we will earn $2.50 in interest, raising our account balance to $1002.50.

In the second month,

$P = \$1002.50$
$I = \$1002.50(0.0025) = \$2.51$ (rounded)
$A = \$1002.50 + \$2.51 = \$1005.01$

Notice that in the second month we earned more interest than we did in the first month. This is because we earned interest not only on the original $1000 we deposited, but we also earned interest on the $2.50 of interest we earned the first month. This is the key advantage that compounding gives us.

Calculating out a few more months in a table or a spreadsheet we have:

Month	Starting balance	Interest earned	Ending Balance
1	1000.00	2.50	1002.50
2	1002.50	2.51	1005.01
3	1005.01	2.51	1007.52
4	1007.52	2.52	1010.04
5	1010.04	2.53	1012.57
6	1012.57	2.53	1015.10
7	1015.10	2.54	1017.64
8	1017.64	2.54	1020.18
9	1020.18	2.55	1022.73
10	1022.73	2.56	1025.29
11	1025.29	2.56	1027.85
12	1027.85	2.57	1030.42

To find an equation to represent this, we will go through a few months to see the pattern:

Initial Amount: $P = \$1000$

1st Month $A = 1.0025(\$1000)$

2nd Month $A = 1.0025(1.0025(\$1000)) = 1.0025^2(\$1000)$

3rd Month $A = 1.0025(1.0025^2(\$1000)) = 1.0025^3(\$1000)$

4th Month $A = 1.0025(1.0025^3(\$1000)) = 1.0025^4(\$1000)$

Observing a pattern, we could conclude

n^{th} month $A = 1.0025^n(\$1000)$

Notice that the \$1000 in the equation was *P*, the starting amount. We found 1.0025 by adding one to the interest rate divided by 12, since we were compounding 12 times per year. Generalizing our result, we could write

Compound Interest

$$A = P\left(1+\frac{r}{n}\right)^{nt} \qquad \text{or} \qquad P = \frac{A}{\left(1+\frac{r}{n}\right)^{nt}}$$

A	is the future value balance in the account after n years
P	is the principal or present value
r	is the annual interest rate in decimal form
n	is the number of compounding periods in one year
t	is the number of years

If the compounding is done annually (once a year), $n = 1$.
If the compounding is done quarterly, $n = 4$.
If the compounding is done monthly, $n = 12$.
If the compounding is done weekly, n = 52
If the compounding is done daily, $n = 365$.

The most important thing to remember about using this formula is that it assumes that we put money in the account **once** and let it sit there earning interest.

The Future Value Spreadsheet Formula

The compound interest formula is built into spreadsheets and is called the future value formula.

Future Value Spreadsheet Formula

=FV(rate per period, number of periods, payment amount, present value)

rate per period	is the interest rate per compounding period, *r/n*
number of periods	is the total number of periods, *n*t*
payment amount	is the amount of regular payments. If none, enter *0*
present value	is the amount deposited or principal, *P*

We will use the payment amount in a future section, for now that will be 0. There is also an optional input at the end to specify making payments at the beginning or end of the period, but we will not use it in this book.

Now let's look at an example and calculate the compound interest using the spreadsheet and the mathematical formula.

Example 4:

A certificate of deposit (CD) is a savings instrument that many banks offer. It usually gives a higher interest rate, but you cannot access your investment for a specified length of time. Suppose you deposit $3000 in a CD paying 6% interest, compounded monthly. How much will you have in the account after 20 years?

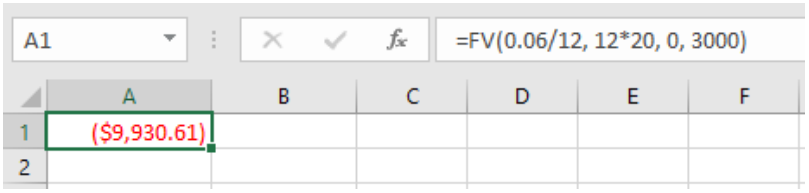
$P = \$3000$	the initial deposit
$r = 0.06$	6% annual rate
$n = 12$	12 months in 1 year
$t = 20$	since we're looking for how much we'll have after 20 years

To use a spreadsheet, you would enter

=FV(rate per period, number of periods, payment amount, present value)

=FV(0.06/12, 12*20, 0, 3000)

=$9,930.61, rounded to the nearest cent.



Note that the output of the formula gives the answer with the opposite sign as the principal and payments. A negative number may be denoted with a negative sign or with the color red or parentheses. The signs may be used in accounting, but we will ignore them in this book.

To use a formula, we are looking for the future value, so we use the formula solved for A:

$$A = 3000\left(1+\frac{0.06}{12}\right)^{12\cdot 20}$$

$$= \$9,930.61$$

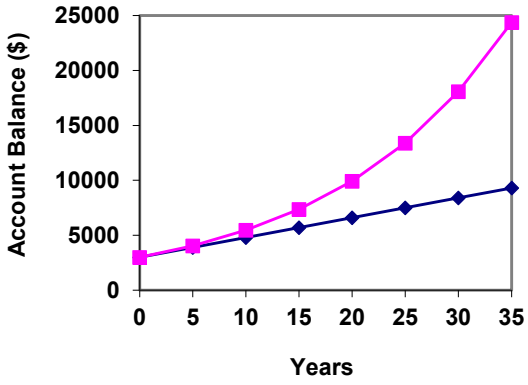
To use a calculator, you would enter the formulas including parentheses around any inside operations. You would enter

$3000\left(1+\left(0.06/12\right)\right)\text{^}\left(12\cdot 20\right)=\$9,930.61.$

Comparing Simple and Compound Interest

Let us compare the amount of money earned from compounding in the previous example against the amount you would earn from simple interest. From the table and graph below we can see that over a long period of time, compounding makes a large difference in the account balance. You may recognize this as the difference between linear growth and exponential growth.

Years	Simple Interest ($15 per month)	Compound Interest (6% compounded monthly or 0.5% each month)
0	$3000	$3000
5	$3900	$4046.55
10	$4800	$5458.19
15	$5700	$7362.28
20	$6600	$9930.61
25	$7500	$13394.91
30	$8400	$18067.73
35	$9300	$24370.65



Finding the Principal, or Present Value

When we know the amount of money we want to have in the future, we can use the formula that is solved for *P*. It requires a little algebra to divide both sides of the formula by the quantity that was multiplied by *P*. There is also a spreadsheet formula which we will introduce now, and then do an example using both methods.

The Present Value Spreadsheet Formula

The present value spreadsheet formula will calculate how much you need to deposit in the present to get a specified future value.

Present Value Spreadsheet Formula

=PV(rate per period, number of periods, payment amount, future value)

rate per period	is the interest rate per compounding period, *r/n*
number of periods	is the total number of periods, *n*t*
payment amount	is the amount of regular payments. If none, enter *0*
future value	is the amount desired in the future, *A*

Example 5:

You know that you will need $40,000 for your child's education in 18 years. If your account earns 4% compounded quarterly, how much would you need to deposit now to reach your goal?

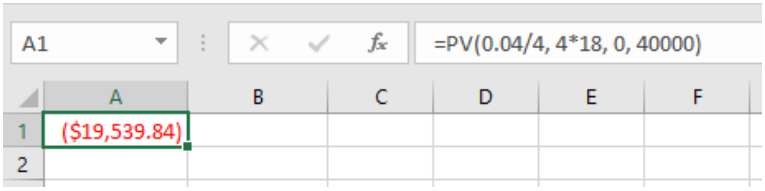
We are looking for what we need to deposit now so we will use the present value formula. We type the formula and inputs the same way we used the future value formula.

=PV(rate per period, number of periods, payment amount, future value)

=PV(0.04/4, 4*18, 0, 40000)

= $19,539.84

You would need to deposit $19,539.84 now and keep the same interest rate to have $40,000 in 18 years.



Note that we cannot enter commas in numbers in a spreadsheet. Commas are used to separate the input values, so we would not get the same answer if we put in $40,000 for an input.

To use the mathematical formula, we use the one that is solved for P.

$r = 0.04$	4%
$n = 4$	4 quarters in 1 year
$t = 18$	Since we know the balance in 18 years
$A = \$40{,}000$	The amount we have in 18 years

In this case, we're going to have to set up the equation, and solve for P.

$$P = \frac{40000}{\left(1+\frac{0.04}{4}\right)^{4\cdot 18}}$$
$$= 19{,}539.84$$

You would need to deposit $19,539.84 now to have $40,000 in 18 years.

Continuously Compounded Interest

In many bank accounts your interest is compounded continuously, or at each moment in time. The number of times per year, n, is infinite. As n approaches infinity the compound interest formula changes to the continuously compounded interest formula.

Continuously Compounded Interest

$$A = Pe^{rt} \quad \text{or} \quad P = \frac{A}{e^{rt}}$$

A is the future value or desired balance in the account
P is the principal or present value
r is the annual interest rate in decimal form
t is the number of years
e is an irrational number that is approximately 2.718281828…
Find e on your calculator to use this formula

To calculate this on a spreadsheet we use the =EXP function. The spreadsheet formulas are

=Principal*EXP(r*t) or =A/EXP(r*t)

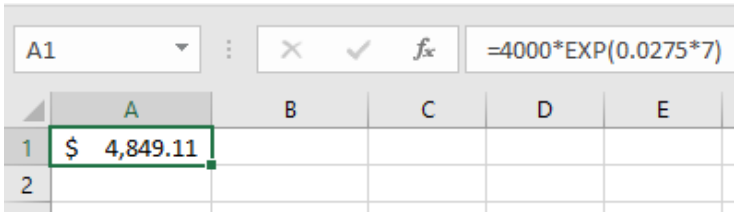
Example 6:

You deposit $4000 in an account that earns 2.75% interest compounded continuously. How much will you have after 7 years? How much interest did you earn? What percentage of the final balance is interest?

To use a spreadsheet, we look at the formula solved for A, the future value. We enter

=4000*EXP(0.0275*7)

=$4849.11.



To use the formula, we have:

$A = Pe^{rt}$

$P = 4000$ Amount invested
$r = .0275$ Interest rate
$t = 7$ Number of years

$$A = 4000e^{(0.0275 \cdot 7)}$$
$$= \$4,849.11$$

After 7 years your account would be worth $4,849.11. Next, we will calculate the amount of interest earned and the percentage.

Finding the Amount of Interest Earned and the Percentage

In the previous example we also want to know how much interest was earned and what percentage of the final balance is from interest. The future value of the investment is $4,849.11. Now to figure out how much of that was interest, we need to subtract the amount initially deposited.

Example 6 Continued:

To find the total amount of interest earned, we subtract the principal from the total balance.

$$\$4,849.11 - \$4,000 = \$849.11$$

The spreadsheet calculation is

=4849.11 – 4000

=$849.11.

You would earn $849.11 in interest.

To find the percentage that is interest, divide the amount of interest by the total amount.

$$\frac{\$849.11}{\$4849.11} = 0.1751 \text{ or } 17.5\%$$

The spreadsheet calculation is the same:

=849.11/4849.11 =0.1751 or 17.5%.

This tells us that after 7 years, 17.5% of the account was earned as interest.

Effective Rate

If you are shopping around for different investments, you might need to compare different rates that have different compounding periods. If the rate and period are different, it's hard to know which account will give the better result. There is a spreadsheet formula called =EFFECT which will allow us to compare accounts. This is also sometimes called the annual percentage yield, or APY.

Effective Rate Formula

=EFFECT(stated rate, number of compounding periods)

stated rate	is the interest rate given (APR)
number of compounding periods	is the number of times the account is compounded per year, n

Example 7:

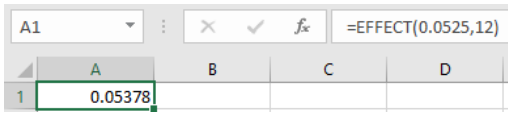
You are comparing an account that pays 5.25% interest compounded monthly, with an account that pays 5% compounded daily. Which account will earn you more interest?

It is hard to tell whether the higher interest rate will be better or the higher compounding rate in this case. We will find the effective rate of both accounts.

For the 5.25% APR account compounded monthly:

=EFFECT(0.0525,12)

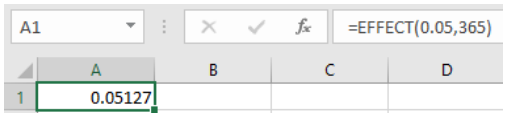
=0.05378 or 5.38%



For the 5% APR account compounded daily:

=EFFECT(0.05,365)

=0.05127 or 5.13%



Now we can compare the effective rates of 5.38% and 5.13% and see that the account with the higher interest rate will earn more interest in this case. This is not always true, so we will show another example.

Example 8:

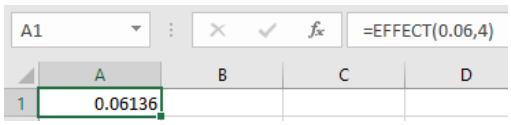
Find the effective rates to compare an account that earns 6% compounded quarterly with an account that earns 5.975% compounded daily. Which one would you choose?

Using the effective rate formula for each, we have:

For the 6% APR account compounded quarterly:

=EFFECT(0.06,4)

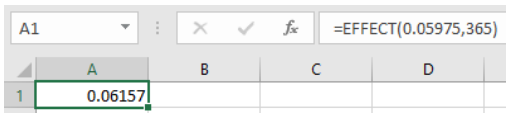
=0.06136 or 6.14%



For the 5.975% APR account compounded daily:

=EFFECT(0.05975,365)

=0.06157 or 6.16%



The account that was compounded more often has a slightly higher rate in this case.

Exercises 2.2

1. A friend lends you $200 for a week, which you agree to repay with 5% one-time interest. How much will you have to repay?

2. You deposit $1,000 in an account that earns simple interest. The annual interest rate is 2.5%.
    a. How much interest will you earn in 5 years?
    b. How much will you have in the account in 5 years?

3. How much will $1,000 deposited in an account earning 7% interest compounded weekly be worth in 20 years?

4. Suppose you obtain a $3,000 Certificate of Deposit (CD) with a 3% annual rate, paid quarterly, with maturity in 5 years.
    a. What is the future value of the CD in 5 years?
    b. How much interest will you earn?
    c. What percent of the balance is interest?

5. You deposit $300 in an account earning 5% interest compounded annually. How much will you have in the account in 10 years?
    a. How much will you have in the account in 10 years?
    b. How much interest will you earn?
    c. What percent of the balance is interest?

6. You deposit $2,000 in an account earning 3% interest compounded monthly.
    a. How much will you have in the account in 20 years?
    b. How much interest will you earn?
    c. What percent of the balance is interest?

7. You deposit $10,000 in an account earning 4% interest compounded weekly.
    a. How much will you have in the account in 25 years?
    b. How much interest will you earn?
    c. What percent of the balance is interest?

8. How much would you need to deposit in an account now in order to have $6,000 in the account in 8 years? Assume the account earns 6% interest compounded monthly.

9. How much would you need to deposit in an account now in order to have $20,000 in the account in 4 years? Assume the account earns 5% interest compounded quarterly.

10. Breylan invests $1,200 in an account that earns 4.6% compounded quarterly and Angad invests the same amount in an account that earns 4.55% compounded weekly.
    a. What will their balances be after 15 years?
    b. What will their balances be after 30 years?
    c. What is the effective rate for each account?

11. Bill invests $6,700 in a savings account that compounds interest monthly at a rate of 3.75%. Ted invests $6,500 in a savings account that compound interest annually at a rate of 3.8%.
    a. Find the effective rate for each account.
    b. Who will have the higher accumulated balance after 5 years?

12. Bassel is comparing two accounts where one pays 3.45% quarterly and the second pays 3.4% daily.
    a. What is the effect rate for each?
    b. If he has $5,000 to deposit how much will the balance be in 10 years?

13. You deposit $2,500 into an account earning 4% interest compounded continuously.
    a. How much will you have in the account in 10 years?
    b. How much total interest will you earn?
    c. What percent of the balance is interest?

14. You deposit $1,000 into an account earning 5.75% compounded continuously.
    a. How much will you have in the account in 15 years?
    b. How much total interest will you earn?
    c. What percent of the balance is interest?

15. You deposit $5,000 in an account earning 4.5% compounded continuously.
    a. How much will you have in the account in 5 years?
    b. How much total interest will you earn?
    c. What percent of the balance is interest?

16. You deposit $10,000 in an account that earns 5.5% compounded continuously and your friend deposits $10,000 in an account that earns 5.5% annually.
    a. How much more will you have in the account in 10 years?
    b. How much more interest did you earn in the 10 years?

Section 2.3 Savings Plans

For most of us it is not practical to deposit a large sum of money in the bank. Instead, we save by depositing smaller amounts of money regularly. We might save in an IRA or 401-K for retirement. We might also save for a down payment on a car or house, or in a college savings plan for our children.

Just like the last section, we will emphasize spreadsheets but calculate each example with the formulas as well. Check with your instructor for which way you should do your problems.

Savings Plan Formulas

To make calculations for savings plans using a spreadsheet, we can use the =FV formula we have already used. This time for regular payments will use the field for payment amount. If we are not making an initial deposit, the present value will be zero.

Here is the future value formula again:

Future Value Spreadsheet Formula

=FV(rate per period, number of periods, payment amount, present value)

rate per period	is the interest rate per compounding period, *r/n*
number of periods	is the total number of periods, *n*t*
payment amount	is the amount of regular payments
present value	is the initial principal. If none, enter 0

The mathematical formulas are shown below. If you want to know how we got the formula, it is derived at the end of the chapter.

Savings Plan Formulas

$$A = \frac{d\left[\left(1+\frac{r}{n}\right)^{nt}-1\right]}{\left(\frac{r}{n}\right)} \quad \text{or} \quad d = \frac{A\left(\frac{r}{n}\right)}{\left[\left(1+\frac{r}{n}\right)^{nt}-1\right]}$$

A	is the balance in the account after n years (future value)
d	is the regular deposit (or **payment amount** each month, quarter, year, etc.)
r	is the annual interest rate in decimal form
n	is the number of compounding periods in one year
t	is the number of years

If the compounding frequency is not explicitly stated, assume there are the same number of compounds in a year as there are deposits made in a year.

If you make your deposits every year, use yearly compounding, $n = 1$.
If you make your deposits every quarter, use quarterly compounding, $n = 4$.
If you make your deposits every month, use monthly compounding, $n = 12$, etc.

To see how both of these methods work, let's look at an example.

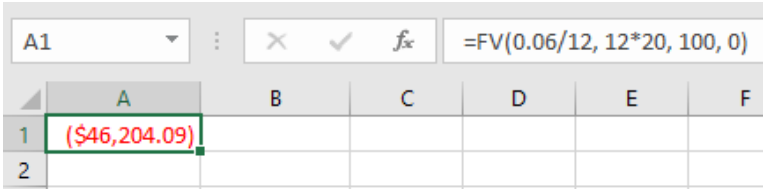
Example 1:

A traditional individual retirement account (IRA) is a special type of retirement account in which the money you invest is exempt from income taxes until you withdraw it. If you deposit $100 each month into an IRA earning 6% interest, how much will you have in the account after 20 years? How much will you have earned in interest? What percentage of the balance is interest?

To use a spreadsheet, we will use =FV because we want to know the balance in the future. We enter 100 for the payment amount and 0 for the present value:

=FV(0.06/12, 12*20, 100, 0)

=$46,204.09



Remember that the output of the formula gives the answer with the opposite sign as the principal and payments. For our purposes we will ignore the signs.

To use the formula, we use the one solved for A, since we want to know the final amount.

$d = \$100$ the monthly deposit
$r = 0.06$ 6% annual rate
$n = 12$ since we're doing monthly deposits, we'll compound monthly
$t = 20$ we want the amount after 20 years

Putting this into the equation we have:

$$A = \frac{100\left[\left(1+\frac{0.06}{12}\right)^{12\cdot 20}-1\right]}{\left(\frac{0.06}{12}\right)}$$

$$= \frac{100\left[(1.005)^{240}-1\right]}{(0.005)}$$

$$= \$46,204.09$$

With U.S. dollars we round to the nearest cent. The account will grow to $46,204.09 after 20 years.

To find the **amount of interest earned**, calculate the total of all your deposits.

$\$100(20)(12) = \$24,000$

The difference between the total amount and the deposits is the interest earned.

$\$46,204.09 - \$24,000 = \$22,204.09$.

The total amount of interest you earned was $22,204.09.

To find the **percentage of the balance that is interest** we will divide the interest by the total balance.

$$\frac{\$22,204.09}{\$46,204.09} = 0.48056 \text{ or } 48.1\%$$

After 20 years 48.1% of the balance is from interest.

Now here's an example with an initial deposit **and** monthly deposits. We can do this with the spreadsheet formula.

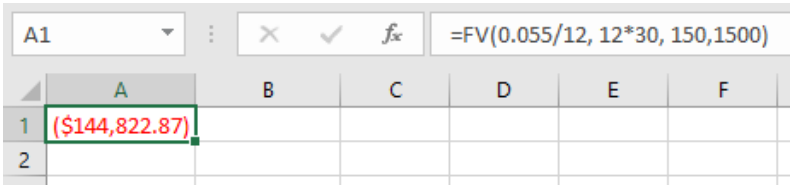
Example 2:

You want to jumpstart your saving by depositing $1500 from your tax return and then deposit $150 every month into an account that earns 5.5% compounded monthly. How much will you have in the account after 30 years?

Using the spreadsheet formula, we can enter an initial deposit and a monthly payment. We enter

=FV(0.055/12, 12*30, 150,1500)

=$144,822.87.



Finding Payment Amounts Spreadsheets and the Formula

Another important thing we can calculate is how much we need to save in each period to have a specified amount in the future. Say you want to achieve a certain amount for retirement or for your kids' college.

The mathematical formula for this is the one solved for *d*, the payment amount, above. There is a new spreadsheet formula to calculate payments, =PMT, that we will introduce now.

Payment Spreadsheet Formula

=PMT(rate per period, number of periods, present value, future value)

rate per period	is the interest rate per compounding period, *r/n*
number of periods	is the total number of periods, *n*t*
present value	is the amount deposited or principal, *P*
future value	is the amount you want in the future, *A*

Here is an example of a retirement goal calculated with a spreadsheet and the formula.

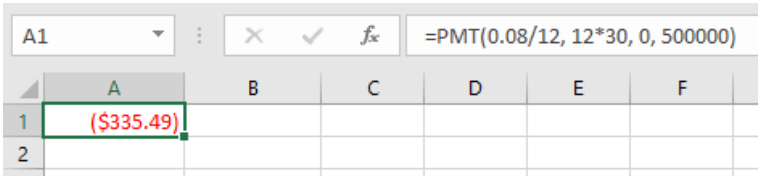
Example 2:

You want to have half a million dollars in your account when you retire in 30 years. Your retirement account earns 8% interest. How much do you need to deposit each month to meet your retirement goal?

To calculate this with a spreadsheet, we will use the =PMT function and enter 0 for the present value and $500,000 for the future value. We cannot enter commas within the numbers however, because spreadsheets use commas to separate the inputs. We enter:

=PMT(0.08/12, 12*30, 0, 500000)

=$335.49.



To see how this works with the formulas, we use the one solved for d, the regular deposit amount.

$r = 0.08$	8% annual rate
$n = 12$	since we're depositing monthly
$t = 30$	30 years
$A = \$500{,}000$	The amount we want to have in 30 years

$$d = \frac{A\left(\frac{r}{n}\right)}{\left[\left(1+\frac{r}{n}\right)^{nt}-1\right]}$$

$$= \frac{500000\left(\frac{0.08}{12}\right)}{\left[\left(1+\frac{0.08}{12}\right)^{12\cdot 30}-1\right]}$$

$$= \$335.49$$

So, you would need to deposit $335.49 each month to have $500,000 in 30 years if your account earns 8% interest.

A note about rounding

If you are using the formulas and round during intermediate steps you will probably have some roundoff error. For this reason, we enter the whole expression into the calculator and do not show the intermediate steps.

One of the challenges in this chapter is choosing the correct formula or spreadsheet function. Read this next example and see if you can determine which formula to use.

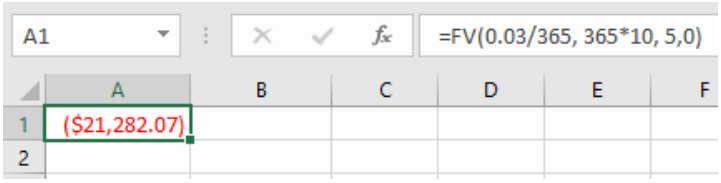
Example 3:

A more conservative investment account pays 3% interest. If you deposit $5 a day into this account, how much will you have after 10 years? What amount and percentage are from interest?

In this example we are given the regular deposit amount and we are looking for the future value. In a spreadsheet we use the =FV function and enter:

=FV(0.03/365, 365*10, 5,0)

=$21,282.07



To use a mathematical formula, we choose the one solved for A:

$d = \$5$ the daily deposit
$r = 0.03$ 3% annual rate
$n = 365$ since we're doing daily deposits, we'll compound daily
$t = 10$ we want the amount after 10 years

$$A = \frac{5\left[\left(1+\frac{0.03}{365}\right)^{365\cdot 10} - 1\right]}{\left(\frac{0.03}{365}\right)}$$

$$= \$21,282.07$$

To find the amount of interest, we will calculate how much was deposited in the account. Since you put in $5 a day for 10 years we get

$$\$5(365)(10) = \$18,250.$$

The interest earned is

$$\$21,282.07 - \$18,250 = \$3,032.07$$

To find the percentage we divide by the total balance to get

$$\frac{\$3032.07}{\$21,282.07} = 0.1424 \text{ or } 14.24\%$$

After 10 years, about 14.2% of the account is interest.

Comparing Lump Sum and Regular Savings Payments

Now let's compare two scenarios to do some multistep problems and get a sense for the value of compounding over time.

Example 4:

Scenario 1: Suppose you invest $200 a month for 15 years into an account earning 10% compounded monthly. After 15 years, you leave the money, without making additional deposits, in the account for another 20 years. How much will you have in the end?

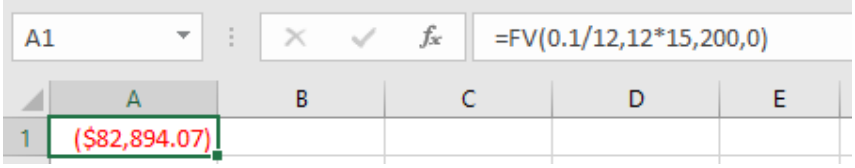
Scenario 2: Suppose instead you didn't invest anything for the first 15 years, then deposited $200 a month for 20 years into an account earning 10% compounded monthly. How much will you have in the end?

Before we calculate the balance for both scenarios, which one do you think will have a higher balance at the end?

For scenario 1, there are two steps involved. The first part is the monthly payments for 15 years. To calculate this with a spreadsheet we enter

=FV(0.10/12, 12*15, 200,0)

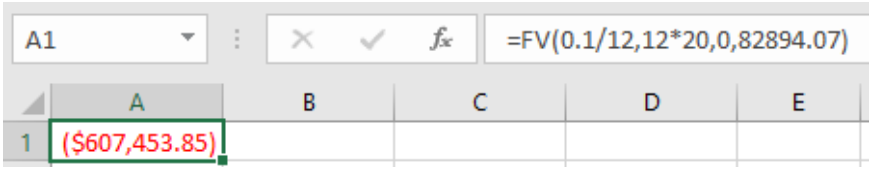
=$82,894.07.



Now you will stop making payments and let the money sit and earn interest for 20 more years. With a spreadsheet we enter

=FV(0.10/12, 12*20, 0, 82894.07)

=$607,453.85.



The process is similar with the formulas. For the first step we have:

$$A = \frac{200\left[\left(1+\frac{0.10}{12}\right)^{12\cdot 15}-1\right]}{\left(\frac{0.10}{12}\right)}$$

$$= \$82,894.07$$

And for the second step we use the compound interest formula from section 2.2.

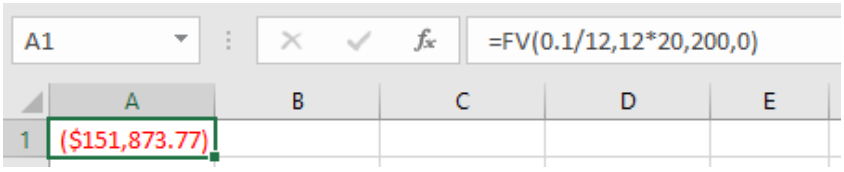
$$A = \$82,894.07\left(1+\frac{0.10}{12}\right)^{12\cdot 20}$$

$$= \$607,453.85$$

Now for Scenario 2: Since we are not investing anything for the first 15 years there is nothing to calculate. This is a one-step problem. We will find the future value with the monthly payments of $200 for 20 years. With a spreadsheet we enter

=FV(0.10/12, 12*20, 200,0)

=$151,873.77.



To check that with the formula we have:

$$A = \frac{200\left[\left(1+\frac{0.10}{12}\right)^{12\cdot 20}-1\right]}{\left(\frac{0.10}{12}\right)}$$

$$= \$151,873.77$$

Were you surprised by these numbers? You would put in less money in scenario 1 and end up with four times as much. The key to compounding interest is to start early. If you remember the graph of compound interest in section 2.2, we can see that as time goes on, the balance increases exponentially.

Deriving the Savings Plan Formula (Optional)

If you are interested in where the savings plan formula came from, we will explain it here. A savings plan with regular payments can be described recursively. Recall that basic compound interest follows from the relationship for each compound period.

$$A = P\left(1+\frac{r}{n}\right)$$

For a savings plan, we need to add a deposit, d, to the account with each compounding period:

$$A = P\left(1+\frac{r}{n}\right)+d$$

Taking this equation from recursive form to explicit form is a bit trickier than with compound interest. It will be easiest to see by working with an example rather than working in general.

Suppose we will deposit $100 each month into an account paying 6% interest. We assume that the account is compounded with the same frequency as we make deposits unless stated otherwise.

In this example:

$r = 0.06$ 6% interest
$n = 12$ 12 compounds/deposits per year
$d = \$100$ our deposit per month

Writing out the recursive equation gives where A is exchanged with P_m where m is the number of compounding periods.

$$P_m = \left(1 + \frac{0.06}{12}\right) P_{m-1} + 100 = (1.005) P_{m-1} + 100$$

Assuming we start with an empty account, we can begin using this relationship:

$$P_0 = 0$$
$$P_1 = (1.005) P_0 + 100 = 100$$
$$P_2 = (1.005) P_1 + 100 = (1.005)(100) + 100 = 100(1.005) + 100$$
$$P_3 = (1.005) P_2 + 100 = (1.005)(100(1.005) + 100) + 100 = 100(1.005)^2 + 100(1.005) + 100$$

Continuing this pattern, after m deposits, we'd have saved:

$$P_m = 100(1.005)^{m-1} + 100(1.005)^{m-2} + \cdots + 100(1.005) + 100$$

In other words, after m months, the first deposit will have earned compound interest for m-1 months. The second deposit will have earned interest for $m-2$ months. Last month's deposit would have earned only one month worth of interest. The most recent deposit will have earned no interest yet.

This equation leaves a lot to be desired, though – it doesn't make calculating the ending balance any easier! To simplify things, multiply both sides of the equation by 1.005:

$$1.005 P_m = 1.005\left(100(1.005)^{m-1} + 100(1.005)^{m-2} + \cdots + 100(1.005) + 100\right)$$

Distributing on the right side of the equation gives

$$1.005 P_m = 100(1.005)^m + 100(1.005)^{m-1} + \cdots + 100(1.005)^2 + 100(1.005)$$

Now we'll line this up with like terms from our original equation, and subtract each side

$$\begin{array}{rcccccc} 1.005 P_m & = & 100(1.005)^m + & 100(1.005)^{m-1} + \cdots + & 100(1.005) & \\ P_m & = & & 100(1.005)^{m-1} + \cdots + & 100(1.005) & +100 \end{array}$$

Almost all the terms cancel on the right side when we subtract, leaving

$$1.005 P_m - P_m = 100(1.005)^m - 100$$

Solving for P_m

$$0.005P_m = 100\left((1.005)^m - 1\right)$$

$$P_m = \frac{100\left((1.005)^m - 1\right)}{0.005}$$

Replacing P_m with A (Future Value), m months with $12t$, where t is measured in years, gives

$$A = \frac{100\left[(1.005)^{12t} - 1\right]}{0.005}$$

Recall 0.005 was r/n and 100 was the deposit d. The value 12 was n, the number of deposits each year. Generalizing this result, we get the savings plan formula solved for A. The second formula uses algebra to rearrange the formula to be solved for d.

Savings Plan Formulas

$$A = \frac{d\left[\left(1+\frac{r}{n}\right)^{nt} - 1\right]}{\left(\frac{r}{n}\right)} \quad \text{or} \quad d = \frac{A\left(\frac{r}{n}\right)}{\left[\left(1+\frac{r}{n}\right)^{nt} - 1\right]}$$

A is the balance in the account after n years (future value)
d is the regular deposit (the amount you deposit each year, each month, etc.)
r is the annual interest rate in decimal form.
n is the number of compounding periods in one year
t is the number of years

If the compounding frequency is not explicitly stated, assume there are the same number of compounds in a year as there are deposits made in a year

Exercises 2.3

1. You set up a savings plan for retirement in 35 years. You will deposit $250 each month for 35 years. The account will earn an average of 6.5% compounded monthly.
    a. How much will you have in your retirement plan in 35 years?
    b. How much interest did you earn.
    c. What percent of the balance is interest?

2. You set up a savings plan for retirement in 40 years. You will deposit $75 each week for 40 years. The account will earn an average of 8.5% compounded weekly
    a. How much will you have in your retirement plan in 40 years?
    b. How much interest did you earn.
    c. What percent of the balance is interest?

3. You set up a savings plan for retirement in 30 years. You will deposit $750 each quarter for 30 years. The account will earn an average of 7.75% compounded quarterly.
    a. How much will you have in your retirement plan in 30 years?
    b. How much interest did you earn?
    c. What percent of the balance is interest?

4. Suppose you invest $130 a month for 5 years into an account earning 9% compounded monthly. After 5 years, you leave the money, without making additional deposits, in the account for another 25 years.
    a. How much will you have in the end?
    b. How much interest did you earn?
    c. What percent of balance is interest?

5. Suppose you have 30 months in which to save $3,500 for a cruise for your family. If you can earn an APR of 3.8%, compounded monthly, how much should you deposit each month?

6. You wish to have $3,000 in 2 years to buy a fancy new stereo system. How much should you deposit each quarter into an account paying 6.5% compounded quarterly?

7. Jamie has determined they need to have $450,000 for retirement in 30 years. Their account earns 6% interest. How much would Jamie need to deposit in the account each month?

8. Lashonda already knows that she wants $500,000 when she retires. If she sets up a saving plan for 40 years in an account paying 10% interest, compounded quarterly how much should she deposit each quarter?

9. Jose' inherits $55,000 and decides to put it in the bank for the next 25 years to save for his retirement. He will earn an average of 5.6% compounded monthly for the next 25 years. His partner deposits $375 a month in a separate savings plan that earns 5.6% interest compounded monthly for the next 25 years.
    a. How much will each have at the end of 25 years?
    b. How much interest did each person earn?
    c. What percent of balance is interest for each person?

10. Akiko inherits $45,000 and decides to put it in the bank for the next 30 years to save for her retirement. She will earn an average of 7.8% compounded monthly for the next 30 years. Her spouse deposits $200 a month in a separate savings plan that earns 7.8% interest compounded monthly for the next 30 years.
    a. How much will each have at the end of 30 years?
    b. How much interest did each person earn?
    c. What percent of balance is interest for each person?

Section 2.4 Loan Payments

In the last section, you learned about savings plans. In this section, you will learn about conventional loans (also called amortized loans or installment loans). Examples include student loans, car loans and home mortgages. These techniques do not apply to payday loans, add-on loans, or other loan types where the interest is calculated up front.

Loan Formulas

In a savings plan, you start with nothing, put money into an account once or on a regular basis, and have a larger balance at the end. Loans work in reverse. You start with a balance owed, make payments and the future value is zero when the loan is paid off.

We will continue to use the same spreadsheet formulas. The ones that are most useful for loans are =PV and =PMT. We will look at how the inputs change for a loan.

Spreadsheet Formulas

=PV(rate per period, number of periods, payment amount, future value)

=PMT(rate per period, number of periods, present value, future value)

rate per period	is the interest rate per compounding period, *r/n*
number of periods	is the total number of periods, *n*t*
payment amount	is the amount of regular payments, *d*
present value	is the amount deposited or principal, *P*
future value	is the amount you want in the future, *0 for a loan*

These two formulas correspond to the formulas below. The formula for loans is derived in a similar way that we did for savings plans, but notice they have negative exponents. The details are omitted here.

Loan Formulas

$$P = \frac{d\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}{\left(\frac{r}{n}\right)} \quad \text{or} \quad d = \frac{P\left(\frac{r}{n}\right)}{\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}$$

P is the balance in the account at the beginning (the principal, or amount of the loan).
d is your loan payment (your monthly payment, annual payment, etc.)
r is the annual interest rate in decimal form
n is the number of compounding periods in one year
t is the length of the loan, in years

Like before, the compounding frequency is not always explicitly given, but is determined by how often you make payments

Example 1:

Teresa wants to buy a car that costs $15,000. She has $3000 saved for the car and plans to finance the rest. She found a 3-year loan at 2.75% APR and a 5-year loan at 4%. How much will her monthly car payment be for each loan and how do these loans compare to each other.

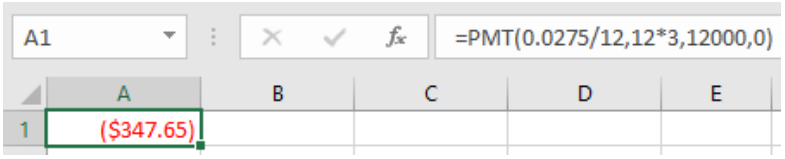
To use a spreadsheet, we use the =PMT formula. For a loan, the loan amount is the present value and the future value is 0, indicating that the loan will be paid off. Teresa is making a down payment, so we also need to subtract that from the cost of the car to find the loan amount:

$15,000 – $3,000 = $12,000

Her loan amount is $12,000. For the 3-year loan at 2.75% APR, we enter:

=PMT(0.0275/12, 12*3, 12000, 0)

=$347.65



For the formula, we use the one solved for *d*:

$r = .0275$ 2.75% annual rate
$n = 12$ monthly payments
$t = 3$ 3 years
$P = 12000$ Since she can pay $3,000 of the $15,000

$$d = \frac{P\left(\frac{r}{n}\right)}{\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}$$

$$= \frac{12000\left(\frac{0.0275}{12}\right)}{\left(1-\left(1+\frac{0.0275}{12}\right)^{-12\cdot 3}\right)}$$

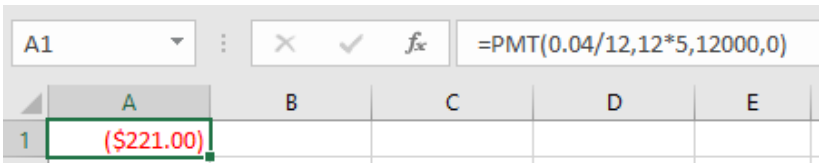
$$= \$347.65$$

Teresa's car payment would be $347.65.

Now for the 5-year loan at 4% APR, we enter:

=PMT(0.04/12, 12*5, 12000, 0)

=$221.00



To use the formula, we have:

$r = .04$ 4% annual rate
$n = 12$ monthly payments
$t = 5$ 5 years
$P = 12000$ the loan amount

$$d = \frac{P\left(\frac{r}{n}\right)}{\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}$$

$$= \frac{12000\left(\frac{0.04}{12}\right)}{\left(1-\left(1+\frac{0.04}{12}\right)^{-12\cdot 5}\right)}$$

$= \$221.00$

Now let's compare the loans by finding out how much Teresa would pay in interest for each loan.

For the 3-year loan at 2.75% APR, her payments would total:

$\$347.65(12)(3) = \$12,515.40$

Her interest would be $515.40.

For the 5-year loan at 4% APR, her payments would total:

$\$221.00(12)(5) = \$13,260.00$

Her interest would be $1,260.00.

There are two main differences between these two loans: the monthly payments and the total paid over the life of the loans. The first loan has a higher monthly payment by $126.65 per month. However, she would pay $744.60 less in interest.

In addition to loan payments, we can calculate the amount of loan we can afford given a monthly payment. Let's look at that in the next example.

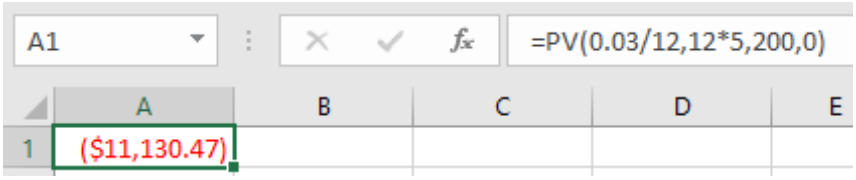
Example 2:

You can afford $200 per month as a car payment. If you can get an auto loan at 3% interest for 60 months (5 years), how expensive of a car can you afford? In other words, what amount loan can you pay off with $200 per month?

To use a spreadsheet for this problem, we use the =PV formula because we want to know the present value, which is the value of the loan right now. We enter

=PV(0.03/12, 12*5, 200,0)

=$11,130.47.



To use a formula, we are looking for P, the starting amount of the loan.

$d = \$200$	the monthly loan payment
$r = 0.03$	3% annual rate
$n = 12$	since we're doing monthly payments, we'll compound monthly
$t = 5$	since we're making monthly payments for 5 years

$$P = \frac{d\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}{\left(\frac{r}{n}\right)}$$

$$= \frac{200\left(1-\left(1+\frac{0.03}{12}\right)^{-12\cdot 5}\right)}{\left(\frac{0.03}{12}\right)}$$

$$= \frac{200\left(1-(1.0025)^{-60}\right)}{(0.0025)}$$

$$= \$11{,}130.47$$

You can afford a maximum loan of $11,130.47. If you have a down payment you can add that to get the value of the car you can buy. If there are any closing costs for the loan you also need to take that into consideration.

To find the amount of interest you will pay for this loan, calculate the total of all your payments.

$$\$200(5)(12) = \$12{,}000$$

Then take the difference between the total payments and the loan amount.

$$\$12{,}000 - \$11{,}130.47 = \$869.53.$$

In this case, you would be paying $869.53 in interest.

So far, we have looked at car loans. Student loans and home mortgages are calculated in the same way. Here is an example of a mortgage payment.

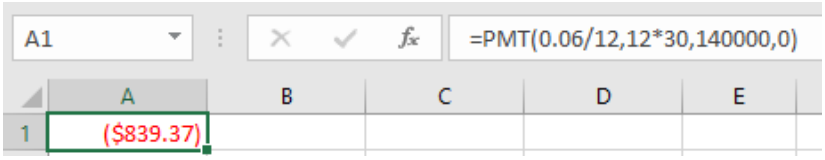
Example 3:

You want to take out a $140,000 mortgage (home loan). The interest rate on the loan is 6%, and the loan is for 30 years. How much will your monthly payments be? What percentage of your total payments will go towards interest?

To use a spreadsheet for this problem, we use the =PMT formula because we want to know the payment amount. The amount of the loan is the present value and to pay off the loan the future value is 0. We enter

=PMT(0.06/12, 12*30, 140000,0)

=$839.37.



To use the formula, we have:

$r = 0.06$	6% annual rate
$n = 12$	since we're paying monthly
$t = 30$	30 years
$P = \$140{,}000$	the starting loan amount

In this case, we're going to use the equation that is solved for *d.*

$$d = \frac{P\left(\frac{r}{n}\right)}{\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}$$

$$= \frac{140000\left(\frac{.06}{12}\right)}{\left(1-\left(1+\frac{.06}{12}\right)^{-12\cdot 30}\right)}$$

$$= \frac{700}{\left(1-(1.005)^{-360}\right)}$$

$$= \$839.37$$

You would make payments of $839.37 per month for 30 years.

To find out what percentage of the total will go towards interest, we need to total up all of the payments.

$$\$839.37(30)(12) = \$302{,}173.20$$

Then take the difference between the total payments and the loan amount.

$\$302{,}173.20 - \$140{,}000 = \$162{,}173.20$.

In this case, you would be paying $162,173.20 in interest over the life of the loan. To find the percentage, we divide the interest by the total amount paid.

$$\frac{\$162{,}173.20}{\$302{,}173.20} = 0.5366 \text{ or } 53.7\%$$

About 53.7% of the total is being paid towards interest.

Remaining Loan Balance

With loans, it is often desirable to determine what the remaining loan balance will be after some number of years. For example, if you purchase a home and plan to sell it in five years, you might want to know how much of the loan balance you will have paid off and how much you will have to pay from the sale.

To determine the remaining loan balance after some number of years, we first need to calculate the payment amount, if we don't already know it. Remember that only a portion of your loan payments go towards the loan balance; a portion is going to go towards interest. For example, if your payments were $1,000 a month, after a year you will *not* have paid off $12,000 of the loan balance.

To determine the remaining loan balance, we can think "how much loan will these loan payments be able to pay off in the remaining time on the loan?"

Example 4:

If a 30-year mortgage at a 6% interest rate has payments of $1,000 a month, what will the loan balance be in 5 years?

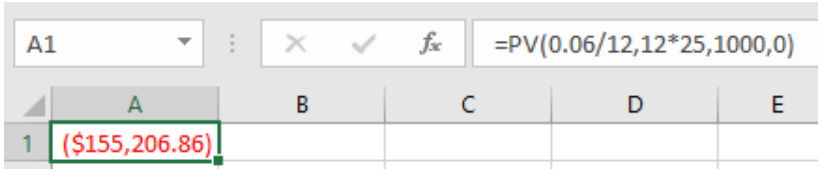
To determine this, we need to think backwards. We are looking for the amount of the loan that can be paid off by $1,000 per month in the remaining 25 years. In other words, we're looking for P when:

$d = \$1{,}000$	the monthly loan payment
$r = 0.06$	6% annual rate
$n = 12$	since we're doing monthly payments, we'll compound monthly
$t = 25$	since we'd be making monthly payments for 25 more years

To use a spreadsheet for this problem, we use the =PV formula because we want to know what the present value would be at the time you want to sell in 5 years. We enter:

=PV(0.06/12, 12*25, 1000,0)

=$155,206.86.



To check this with the formula we have:

$$P = \frac{1000\left(1-\left(1+\frac{0.06}{12}\right)^{-12\cdot 25}\right)}{\left(\frac{0.06}{12}\right)}$$

$$= \frac{1000\left(1-(1.005)^{-300}\right)}{(0.005)}$$

$$= \$155,206.86$$

The loan balance with 25 years remaining on the loan will be $155,206.86

Sometimes answering remaining balance questions requires two steps, both of which we have done in this section:

1. Calculate the monthly payment on the loan
2. Calculate the remaining loan balance based on the *remaining time* on the loan

On the next page we will give a summary of all the spreadsheet formulas we have used and when to use them.

Summary of Spreadsheet Formulas

Here are all the spreadsheet formulas from this chapter so far together so you can see the similarities and differences.

Spreadsheet Formulas

=principal+principal*rate*time

=FV(rate per period, number of periods, payment amount, present value)

=principal*EXP(yearly rate*years)

=PV(rate per period, number of periods, payment amount, future value)

=PMT(rate per period, number of periods, present value, future value)

=EFFECT(stated rate, number of compounding periods per year)

rate per period	is the interest rate per compounding period, *r/n*
number of periods	is the total number of periods, *n*t*
payment amount	is the amount of regular payments, *d*
present value	is the amount deposited or principal, *P*
future value	is the amount you want in the future, *0 for a loan*

When to use the formulas: What is the question asking?

- Find a payment: =PMT
- Find the effective rate or compare accounts: =EFFECT
- How much do you need to deposit now, what loan amount can you afford, or remaining loan balance: =PV

- What will the account balance be in the future?
  - Simple interest: =principal+principal*rate*time
  - Compound interest (except continuous): =FV
  - Continuously compounded interest: principal*EXP(rate*years)

Summary of Mathematical Formulas

Mathematical Formulas

Simple Interest $I = Prt$ $A = P + Prt$

Compound Interest $A = P\left(1+\frac{r}{n}\right)^{nt}$ or $P = \frac{A}{\left(1+\frac{r}{n}\right)^{nt}}$

Continuously Compounded

$A = Pe^{rt}$ or $P = \frac{A}{e^{rt}}$

Savings Plans $A = \frac{d\left[\left(1+\frac{r}{n}\right)^{nt}-1\right]}{\left(\frac{r}{n}\right)}$ or $d = \frac{A\left(\frac{r}{n}\right)}{\left[\left(1+\frac{r}{n}\right)^{nt}-1\right]}$

Loans $P = \frac{d\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}{\left(\frac{r}{n}\right)}$ or $d = \frac{P\left(\frac{r}{n}\right)}{\left(1-\left(1+\frac{r}{n}\right)^{-nt}\right)}$

P is the principal, starting amount, or present value
d is your loan payment (your monthly payment, annual payment, etc.)
r is the annual interest rate in decimal form
n is the number of compounding periods in one year
t is the length of the loan, in years
A is the end amount or future value

If the compounding frequency is not always explicitly given, it is determined by how often you make payments

When to use the formulas: What is the question asking?

- Find a payment
  - Savings payment: savings plan equation (positive exponent) solved for d
  - Loan payment: loan equation (negative exponent) solved for d

- How much do you need to deposit now?
  - Compound interest (except continuous): compound interest formulas solved for P
  - Continuously compounded: the formula with e solved for P
- What loan amount can you afford, or remaining loan balance: loan formula solved for P
- What will the account balance be in the future?
  - One-time deposit:
    - Simple interest: simple interest formula
    - Compound interest (except continuous): compound interest formula solved for A
    - Continuously compounded interest: the formula with e in it, solved for A
  - Regular payments: Savings plan formula solved for A

Remember, the most important part of answering any kind of question, money or otherwise, is first to correctly identify what the question is really asking, and to determine what approach will best allow you to solve the problem. After practicing with the exercises from this section, you can test yourself with the cumulative chapter 2 exercises.

Exercises 2.4

1. You can afford a $700 per month mortgage payment. You've found a 30-year loan at 5.5% interest.
   a. How big of a loan can you afford?
   b. How much total money will you pay the loan company?
   c. How much of that money is interest?

2. Marie can afford a $250 per month car payment. She's found a 5-year loan at 7% interest.
   a. How expensive of a car can she afford?
   b. How much total money will she pay the loan company?
   c. How much of that money is interest?

3. You want to buy a $25,000 car. The company is offering a 2% interest rate for 48 months (4 years). What will your monthly payments be?

4. You decide to finance a $12,000 car at 3% compounded monthly for 4 years. What will your monthly payments be? How much interest will you pay over the life of the loan?

5. You want to buy a $200,000 home. You plan to pay 10% as a down payment and take out a 30-year loan for the rest.
    a. How much is the loan amount going to be?
    b. What will your monthly payments be if the interest rate is 5%?
    c. What will your monthly payments be if the interest rate is 6%?

6. Lynn bought a $300,000 house, paying 10% down, and financing the rest at 6.5% interest for 30 years.
    a. Find her monthly payments.
    b. How much interest will she pay over the life of the loan?
    c. What percentage of your total payment was interest?

7. Emile bought a car for $24,000 three years ago. The loan had a 5-year term at 3% interest rate. How much does he still owe on the car?

8. A friend bought a house 15 years ago, taking out a $120,000 mortgage at 6% for 30 years. How much does she still owe on the mortgage?

Exploration 2.4

1. Pay day loans are short term loans that you take out against future paychecks: The company advances you money against a future paycheck. Either visit a pay day loan company or look one up online. Be forewarned that many companies do not make their fees obvious, so you might need to do some digging or look at several companies.
    a. Explain the general method by which the loan works.
    b. We will assume that we need to borrow $500 and that we will pay back the loan in 14 days. Determine the total amount that you would need to pay back and the effective loan rate. The effective loan rate is the percentage of the original loan amount that you pay back. It is not the same as the APR (annual rate) that is probably published.
    c. If you cannot pay back the loan after 14 days, you will need to get an extension for another 14 days. Determine the fees for an extension, determine the total amount you will be paying for the now 28-day loan, and compute the effective loan rate.

2. Suppose that 10 years ago you bought a home for $110,000, paying 10% as a down payment, and financing the rest at 9% interest for 30 years.
    a. Let's consider your existing mortgage:
        i. How much money did you pay as your down payment?

   ii. How much money was your mortgage (loan) for?
   iii. What is your current monthly payment?
   iv. How much total interest will you pay over the life of the loan?

b. This year, you check your loan balance. Only part of your payments have been going to pay down the loan; the rest has been going towards interest. You see that you still have $88,536 left to pay on your loan. Your house is now valued at $150,000.
   i. How much of the loan have you paid off? (i.e., how much have you reduced the loan balance by? Keep in mind that interest is charged each month - it's not part of the loan balance.)
   ii. How much money have you paid to the loan company so far?
   iii. How much interest have you paid so far?
   iv. How much equity do you have in your home (equity is value minus remaining debt)?

c. Since interest rates have dropped, you consider refinancing your mortgage at a lower 6% rate.
   i. If you took out a new 30-year mortgage at 6% for your remaining loan balance, what would your new monthly payments be?
   ii. How much interest will you pay over the life of the new loan?

d. Notice that if you refinance, you are going to be making payments on your home for another 30 years. In addition to the 10 years you've already been paying, that's 40 years total.
   i. How much will you save each month because of the lower monthly payment?
   ii. How much total interest will you be paying?
   iii. Does it make sense to refinance? (there isn't a correct answer to this question. Just give your opinion and your reason)

Cumulative Chapter 2 Exercises

For each of the following scenarios, determine which formula to use and solve the problem.

1. Keisha received an inheritance of $20,000 and invested it at 6.9% interest, compounded continuously. How much will she have for college in 8 years?

2. Paul wants to buy a new car. Rather than take out a loan, he decides to save $200 a month in an account earning 3.5% interest compounded monthly. How much will he have saved up after 3 years?

3. Sol is managing investments for a non-profit company. They want to invest some money in an account earning 5% interest compounded annually with the goal to have $30,000 in the account in 6 years. How much should Sol deposit into the account?

4. Miao is going to finance new office equipment at a 2.8% rate over a 4-year term. If she can afford monthly payments of $100, how much new equipment can she buy?

5. How much would you need to save every month in an account earning 4.1% interest to have $5,000 saved up in two years.

6. Terry and Jess are buying a house for $405,000 and they can afford to put 10% down. Their interest rate is 4.3% for 30 years. What will their monthly mortgage payment be?

7. You loan your sister $500 for two years and she agrees to pay you back with 3% simple interest per year. How much will she pay you back

8. Zahid starts saving $150 per month in an account that pays 4.8% compounded monthly. If he continues for 20 years, how much will he have? If he waited 10 years instead and put in $300 per month for 10 years with the same interest, how much would he have?

Section 2.5 Income Taxes

A Very Brief History of Taxes

Although Benjamin Franklin famously claimed in 1789 that "in this world nothing can be said to be certain, except death and taxes", it wasn't until 1913 that the 16th amendment was ratified, and income tax was legalized. Prior to this, taxes were primarily collected through tariffs on imported goods, poll taxes, and property taxes. A **poll tax**, also referred to as a head tax, was a fixed amount every liable individual had to pay. Payment of the poll tax was often required before a person could register to vote or be issued a hunting or fishing license.

Tax policy in the United States is a politically divisive issue. In general, Democrats seek to lower taxes for low and middle income earners and raise taxes for the wealthy, while Republicans support a variety of tax cuts and limiting the amount wealthy earners are taxed. There is even debate as to the current length of the tax code! Some claim that the code is over 70,000 pages while others insist it is just over 2,000. Nevertheless, there is at least one thing everyone can agree with - you don't want to be on the wrong side of the Internal Revenue Service (IRS)!

Types of Income Tax

Income tax is a tax that is levied on earned income or profit. Income taxes are collected by the federal government, states, and even some municipalities. Income taxes are an important source of funding, and are used to finance social programs, maintain and expand infrastructure, and provide foreign aid, among other things.

Federal Income Tax

Anyone who earns income over a certain amount (approximately $10,000 for an individual) must file a federal tax return and have taxes collected on their behalf. The amount of federal tax owed is determined by your **filing status** and **taxable income**.

State Income Tax

Forty-one states and the District of Columbia collect taxes on income from wages and investment. Seven states - Washington, Nevada, South Dakota, Wyoming, Alaska, Texas, and Florida – do not collect tax on wage or investment income, and two states – New Hampshire and Tennessee – only collect tax on investment income. The amount of tax collected varies by state, with each state averaging between $500 and $3000 per person.

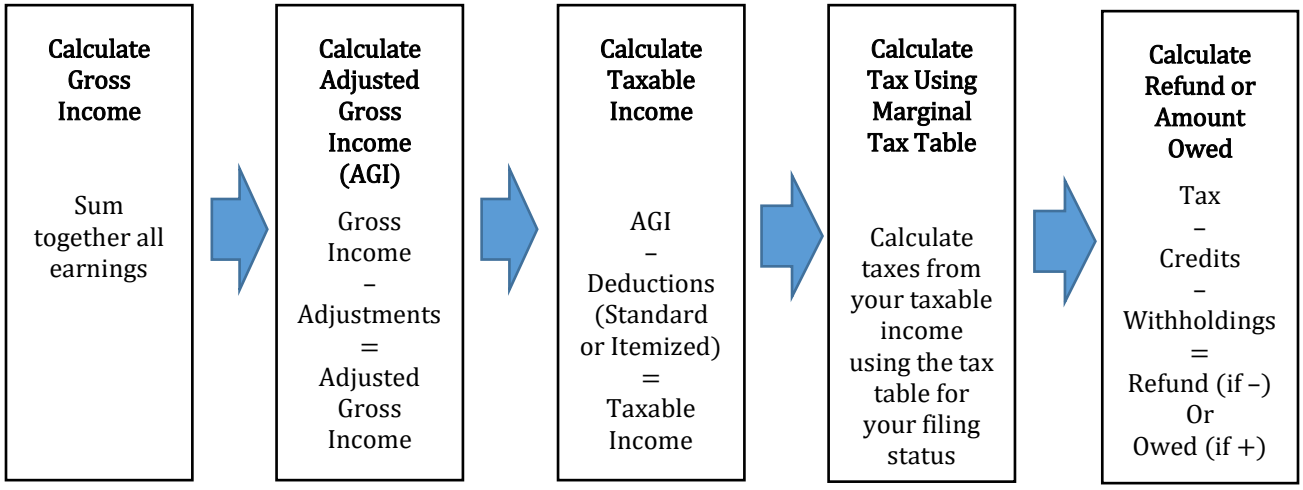
Municipal Income Tax

Some municipalities (urban districts) also collect income taxes. For example, business owners in Multnomah County pay a municipal tax on their business income, and workers in the Tri-met district (area served by Tri-met) have their wages taxed at 0.7537%.

Calculating Federal Income Tax

Not all income is taxed, and not all income that gets taxed is taxed at the same rate. The general process of calculating the amount of federal income tax you owe is outlined in the chart below.

Steps to Calculate Federal Income Tax



Gross Income

Gross income includes all wages, tips, earned interest, dividends, rents and royalties, alimony, property gains, income tax refunds, etc. Keep in mind that ALL income means ALL income. Even income earned from a crime must be reported!

Adjusted Gross Income

Adjustments are eligible expenses used to reduce your gross income. Adjustments are tax-exempt and thus reduce the amount owed in taxes. Eligible expenses include contributions to tax deferred savings plans (401k, Individual Retirement Plan (IRA)), school tuition, student loan interest, moving expenses, business expenses, flexible spending accounts and health savings account contributions.

Taxable Income

Taxable income is determined by subtracting your **deductions** from your adjusted gross income. You may choose to take *either* a **standard deduction**, which is determined by your filing status, or you may choose to **itemized** your deductions. Itemized deductions may include state and local income taxes, property taxes, medical expenses, mortgage interest, and charitable donations.

Prior to the 2018 tax year, you could also claim **personal exemptions**. Exemptions of $4,050 for each member of the household were subtracted from the adjusted gross income along with either the standard or itemized deduction to determine taxable income. Starting in 2018, however, the standard deduction was doubled, and personal exemptions were eliminated.

Let's look at an example of how to calculate gross income and adjusted gross income (AGI).

> Example 1: Sasha received $45,000 in wages and earned $1,300 in interest from his savings plan. He paid $1,400 in student loan interest, and put $4,000 into his Individual Retirement Account (IRA). Determine Sasha's adjusted gross income.
>
> To calculate Sasha's adjusted gross income, we need to first determine his gross income. His gross income includes his wages and earned interest:
>
> Gross Income = $45,000 + $1,300 = $46,300
>
> To find his adjusted gross income, we need to subtract eligible adjustments from his gross income. His eligible adjustments include the interest paid on his student loans, and his contributions to his IRA:
>
> Adjusted Gross Income = $46,300 - $1,400 - $4,000 = $40,900

Filing Status

Your **filing status** is determined by your family situation. Are you married, widowed, divorced, caring for a family member? It is possible to fall into more than one category, but you may choose the one that is most beneficial for you.

> **Single:** If you are unmarried, legally separated from your spouse, or divorced on the last day of the year.
>
> **Married Filing Jointly:** If you are married and both you and your spouse agree to file a joint return. (On a joint return, you report your combined income and deduct your combined allowable expenses.)
>
> **Married Filing Separately:** If you are married and you want to be responsible only for your own tax, or if this status results in less tax than a joint return.
>
> **Head of Household (with qualifying person)**: If you are 1) unmarried or considered unmarried on the last day of the year, 2) paid more than half the cost of keeping up a home for the year, and 3) have a qualifying dependent living with you for more than half the year (except temporary absences, such as school).
>
> **Qualifying Widow/Widower (with dependent child):** This status is available for the two years following the death of your spouse.

Your filing status determines your standard deduction and tax liability. The standard deduction for each filing status for the tax year 2018 is given in the table below.

Filing Status	Standard Deduction for Tax Year 2018
Single	$12,000
Married Filing Jointly	$24,000
Married Filing Separately	$12,000
Head of Household	$18,000
Qualifying Widower	$24,000

Here is an example where we calculate adjusted gross income and taxable income.

> Example 2: Maria earned wages of $32,400 from her job as a server. She also earned $8,300 in tips, and received $95 in interest from a savings account. In trying to save for retirement, Maria has put $3,500 into a 401K tax deferred savings plan. She is unmarried and lives with her six year old daughter. Determine Maria's gross income, adjusted gross income, and taxable income.
>
> Maria's gross income includes her wages, tips, and earned interest.
>
> Gross Income = $32,400 + $8,300 + $95 = $40,795
>
> Maria can use her 401k contribution as an adjustment.
>
> Adjusted Gross Income = $40,795 - $3,500 = $37,295
>
> Since Maria is unmarried and is the primary caregiver for her daughter, she can file as the head of household and take the standard deduction of $18,000.
>
> Taxable Income = $37,295 - $18,000 = $19,295

Tax Tables

The United States has a **progressive** federal income tax system which means the more income you have, the more you generally pay in taxes. There are **marginal tax rates** that go up according to our income, but we don't pay the same rate on all of our income. The ranges of income are called **tax brackets** and the buckets in the illustration below can help us visualize the brackets.

The cutoff values in the figure are for a married couple filing jointly, but the percentages are the same for everyone as you'll see in the table on the next page. We pay 10% of any taxable income in the first bucket. If the first bucket is full, we move to the second bucket and we pay 12% of that income. This couple is in the 22% tax bracket because the 3rd bucket is partially filled. They will pay 22% of the income in that bucket, but they are not paying 22% of their entire income.

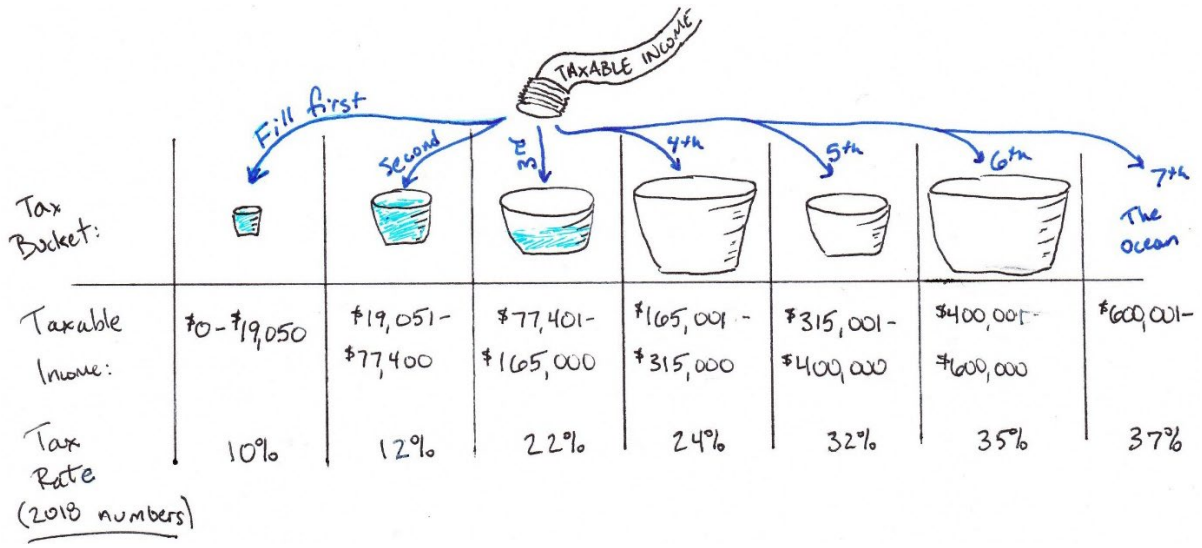


Figure 1. Tax Buckets by John Chesbrough, licensed under CC-BY-ND-NC 4.0.

Here is a table with cutoff values for each filing status. Everyone pays a tax of 10% on any income in the first tax bracket. Then we all pay 12% of the next bracket of income, 22% of the next bracket, and so on. This keeps going up to the 37% tax bracket.

2018 Tax Year				
Marginal Tax Rate On Taxable Income	Filing Single	Filing as Head of Household	Married filing Jointly	Married filing Separately
10%	First $9,525	First $13,600	First $19,050	First $9,525
12%	$9,525 – $38,700	$13,600 – $51,800	$19,050 – $77,400	$9,525 – $38,700
22%	$38,700 – $82,500	$51,800 – $82,500	$77,400 – $165,000	$38,700 – $82,500
24%	$82,500 – $157,500	$82,500 – $157,500	$165,000 – $315,000	$82,500 – $157,500
32%	$157,500 – $200,000	$157,500 – $200,000	$315,000 – $400,000	$157,500 – $200,000
35%	$200,000 – $500,000	$200,000 – $500,000	$400,000 – $600,000	$200,000 – $300,000
37%	Over $500,000	Over $500,000	Over $600,000	Over $300,000

Let's see how to use this table to calculate taxes in an example.

Example 3:

If Avery is filing single and has $55,100 in taxable income, calculate their tax.

We begin with the lowest tax bracket and take 10% of $9,525. Since their income is higher than that, we add 12% of the next amount, found by subtracting the values in that bracket. Avery's taxable income is in the 22% tax bracket, so we find the amount over $38,700 by subtracting. Here is the full calculation:

$$\begin{aligned}&0.10(\$9,525)+0.12(\$38,700-\$9,525)+0.22(\$55,100-\$38,700)\\&=0.10(\$9,525)+0.12(\$29,175)+0.22(\$16,400)\\&=\$952.50+\$3,501+\$3,608\\&=\$4,435.50+\$3,608\\&=\$8,043.50\end{aligned}$$

Avery would owe $8,043.50 in taxes. Note that their overall tax rate is somewhere between 10% and 22%. We can calculate the actual rate by dividing their tax owed by their taxable income:

$$\frac{\$8,043.50}{\$55,100}=0.1459 \text{ or about } 15\%$$

We could calculate all taxes this way, but you might notice that the first few terms will be the same if the buckets are full. For this reason, we can simplify these tax tables. The tax tables below give the value for all of the lower tax buckets that are full. There is a tax table for each filing status and the cutoffs are regularly adjusted for inflation, so they usually vary from year to year.

2018 Federal Income Tax Tables

Single (2018)	
If taxable income is:	The tax is:
Not over $9,525	10% of the taxable income
Over $9,525 but not over $38,700	$952.50 plus 12% of the excess over $9,525
Over $38,700 but not over $82,500	$4,453.50 plus 22% of the excess over $38,700
Over $82,500 but not over $157,500	$14,089.50 plus 24% of the excess over $82,500
Over $157,500 but not over $200,000	$32,089.50 plus 32% of the excess over $157,500
Over $200,000 but not over $500,000	$45,689.50 plus 35% of the excess over $200,000
Over $500,000	$150,689.50 plus 37% of the excess over $500,000

Head of Household (2018)	
If taxable income is:	The tax is:
Not over $13,600	10% of the taxable income
Over $13,600 but not over $51,800	$1,360 plus 12% of the excess over $13,600
Over $51,800 but not over $82,500	$5,944 plus 22% of the excess over $51,800
Over $82,500 but not over $157,500	$12,698 plus 24% of the excess over $82,500
Over $157,500 but not over $200,000	$30,698 plus 32% of the excess over $157,500
Over $200,000 but not over $500,000	$44,298 plus 35% of the excess over $200,000
Over $500,000	$149,298 plus 37% of the excess over $500,000

Married Filing Jointly (2018)	
If taxable income is:	The tax is:
Not over $19,050	10% of the taxable income
Over $19,050 but not over $77,400	$1,905 plus 12% of the excess over $19,050
Over $77,400 but not over $165,000	$8,907 plus 22% of the excess over $77,400
Over $165,000 but not over $315,000	$28,179 plus 24% of the excess over $165,000
Over $315,000 but not over $400,000	$64,179 plus 32% of the excess over $315,000
Over $400,000 but not over $600,000	$91,379 plus 35% of the excess over $400,000
Over $600,000	$161,379 plus 37% of the excess over $600,000

Married Filing Separately (2018)	
If taxable income is:	The tax is:
Not over $9,525	10% of the taxable income
Over $9,525 but not over $38,700	$952.50 plus 12% of the excess over $9,525
Over $38,700 but not over $82,500	$4,453.50 plus 22% of the excess over $38,700
Over $82,500 but not over $157,500	$14,089.50 plus 24% of the excess over $82,500
Over $157,500 but not over $200,000	$32,089.50 plus 32% of the excess over $157,500
Over $200,000 but not over $300,000	$45,689.50 plus 35% of the excess over $200,000
Over $300,000	$80,689.50 plus 37% of the excess over $300,000

Example 4: Suppose Adira is filing as head of household and has a taxable income of $86,450. Calculate her taxes using the individual tax brackets and with the simplified table.

We will calculate Adira's taxes first the long way:

$$0.10(\$13,600)+0.12(\$51,800-\$13,600)+0.22(\$82,500-\$51,800)+0.24(\$86450-\$82,500)$$
$$=0.10(\$13,600)+0.12(\$38,200)+0.22(\$30,700)+0.24(\$3,950)$$
$$=\$1,360+\$4,584+\$6,754+\$948$$
$$=\$12,698+\$948$$
$$=\$13,646$$

Now, using the simplified tax table for single filing status, we see that Adira's taxable income puts her in the 24% tax bracket. The simplified table tells us that her taxes will be equal to $12,698 plus 24% of the excess over $82,500. Notice the number $12,698 is the same value we got for all of the "full" tax brackets, so we only need to calculate the last one. To find the excess over $82,500, we subtract $82,500 from her taxable income. Thus, her taxes are:

$$\begin{aligned}\$12,698+0.24(\$86,450-\$82,500)&=\$12,698+0.24(\$3,950)\\&=\$12,698+\$948\\&=\$13,646\end{aligned}$$

We see that both methods result in the same value, $13,646, for Adira's taxes.

Example 5: Phyllis and Gladys are married and filing jointly. Together their taxable income is $112,000. Use the simplified 2018 tax tables from this section to determine how much they owe in taxes.

Since Phyllis and Gladys are married and filing jointly, their taxable income puts them in the 22% tax bracket. Using the simplified 2018 tax table, their taxes are $8,907 plus 22% of the excess over $77,400:

$$\begin{aligned}\$8,907+0.22(\$112,000-\$77,400) &= \$8,907+0.22(\$34,600)\\ &= \$8,907+\$7,612\\ &= \$16,519\end{aligned}$$

Phyllis and Gladys owe $16,519 in taxes.

Tax Credits

Tax credits are different from deductions in that they reduce the amount of tax you owe by the full amount of the credit, not just a percentage. This makes credits much more valuable than deductions.

Common tax credits include child tax credits, earned income credits, child and dependent care credits, American opportunity tax credits, lifetime learning credits, and various federal energy credits.

Example 6: Shiro is in the 22% tax bracket and itemizes his deductions. How much will his tax bill be reduced if he makes a $1,000 contribution to charity? How much will his bill be reduced if he gets a $1,000 tax credit?

The tax credit will reduce his tax bill by the full amount of the credit, so the $1,000 tax credit will reduce his tax bill by $1,000. As a deduction, however, his contribution to charity will only reduce his tax bill by 22% of the $1,000, or $0.22(\$1{,}000) = \$220$. Tax credits are always better than deductions.

Calculating a Refund or Payment Due

Employers are required to take out an estimated amount for taxes from each of our paychecks. These **withholdings** are taken out, so we don't all have huge tax bills at the end of the year, and so the government has the income it needs to run. When you file your taxes, you compare the amount of tax you actually owe, with the withholdings from your paycheck. If you had more withheld during the year than you owe, you will file for a refund. If your withholdings were less, though, you must pay the difference. Let's look at a couple of examples with tax credits and withholdings.

Example 7: John's taxes are $5,342.50. He can claim an American opportunity tax credit of $2,300 and he had $4,135 withheld from his paychecks. Determine if John will owe money or get a refund.

The first step is to reduce John's taxes by the full amount of the tax credit.

$$\$5,342.50-\$2,300=\$3,042.50$$

Next we subtract the amount withheld from his paychecks by his employer. Since his withholdings are greater than the amount he owes after applying the tax credit, he will receive a refund equal to the difference.

$$\$3042.50-\$4,135=-\$1,092.50$$

John will receive a refund of $1,092.50. Notice that having a negative amount after subtracting credits and withholdings means you will receive a refund, while having a positive amount means you still owe money.

Example 8: As we discovered in earlier, Phyllis and Gladys owe $16,519 in taxes. Their employers withheld $8,980 and they received a $7,500 credit for their purchase of an electric car. Will they receive a refund, or will they need to make a payment?

To determine whether they will receive a refund or if they will still owe money, we will first subtract the full amount of the electric car credit.

$\$16,519 - \$7,500 = \$9,019$

We now subtract their withholdings:

$\$9,019 - \$8,980 = \$39$

Since the value is positive, Phyllis and Gladys still owe $39 in taxes.

Exercises

1. Which decreases your tax bill more a credit or a deduction?

2. Can you take the standard deduction and itemize your deductions?

3. Can you make adjustments to your income and take the standard deduction?

4. If you decide to take the standard deduction what have you considered?

5. If you are married do you have to file your taxes together?

6. Fredrick is concerned about the effect of a raise on his taxes. He's getting a raise of $3,000, putting him into a higher tax bracket by $1,000 dollars. He's concerned about his entire income being taxed at 22% instead of 12%. Should he be concerned? Why or why not?

7. Using the simplified 2018 tax table, determine the income tax owed for a single person who has $80,000 of taxable income.

8. Determine the amount of taxes owed or the refund that would result in this situation:
   - Filing Status: Married filing jointly
   - Gross Income: $125,000
   - Adjustments: $5,600
   - Itemized Deductions: $11,400
   - Credits: $15,000
     a. What is their adjusted gross income (AGI)?
     b. Should they itemize or take the standard deduction?
     c. Use the simplified 2018 tax tables to determine their taxes.
     d. What is their final tax refund or amount still owed?

9. Francis and Edward are planning to get married and they want to determine whether there is an advantage or disadvantage to marrying before the end of the year and filing their taxes jointly. Use the information below to calculate the amount they would owe or receive if they each filed as single, and the amount they would owe or receive if they filed jointly as a married couple. Which is the better choice?
    - Filing Status: TBD
    - Francis' Gross Income: $35,000
    - Edward's Gross Income: $40,000
    - Francis' Adjustments: $7000
    - Edward's Adjustments: $3000
    - Francis' Withholdings: $14000
    - Edward's Withholdings: $5500
    - Francis' Credits: $4000
    - Edward's Credits: $5000

10. Janice is unmarried and has two kids. She earned $76,000 in wages last year, received $750 in interest from a savings account, and contributed $25,000 to a tax deferred savings account. Her itemized deductions are $19,600.
    a. Determine Janice’s gross income.
    b. Determine Janice’s adjusted gross income.
    c. Should Janice take the standard deduction or itemize? Explain.
    d. Determine Janice’s taxable income.
    e. If Janice has $4,200 in child tax credits and had $6,300 withheld for taxes from her wages, will Janice owe money, or will she receive a refund? Calculate the amount.

Chapter 3: Statistics

Student Outcomes for this Chapter

Section 3.1: Overview of the Statistical Process

Students will be able to:

- Define and identify the population, parameter, sample and statistic
- Identify four sampling methods: simple random sample (SRS), stratified, systematic and convenience
- Identify and discuss types of bias association with sampling
- Distinguish between experimental and observational studies
- Explain margin of error and confidence intervals

Section 3.2: Describing Data

Students will be able to:

- Define and identify categorical and quantitative data
- Read and construct frequency tables and relative frequency tables
- Make bar charts and pie charts for categorical variables by hand and/or using technology
- Identify elements of misleading graphs: 3-dimensional graphs, perceptual distortion, misleading scales, stacked bar graphs
- Make histograms for quantitative variables by hand and/or using technology
- Identify the number of modes in a distribution and whether it is symmetric, skewed to the left, or skewed to the right

Section 3.3: Summary Statistics: Measures of Center

Students will be able to:

- Calculate and describe the measures of center: mean and median
- Analyze the relationship of the mean and median to the shape of the data

Section 3.4: Summary Statistics: Measures of Variation

Students will be able to:

- Calculate and describe the measures of variation: standard deviation, range and interquartile range (IQR)
- Calculate the 5-number summary and construct boxplots by hand and/or using technology (boxplots using technology may be modified or not)
- Compare distributions with side-by-side boxplots and percentiles
- Calculate and apply Z-scores

Section 3.1 Overview of the Statistical Process

Introduction to Statistics

We are bombarded by information and statistics every day. But if we cannot distinguish credible information from misleading information, then we are vulnerable to manipulation and making decisions that are not in our best interest. Statistics provides tools for us to evaluate information critically. In this sense, statistics is one of the most important things to know about.

Statistics are often presented to add credibility to an argument. To give some examples, here are some claims that we have heard on several occasions. (We are *not* saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentyne.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

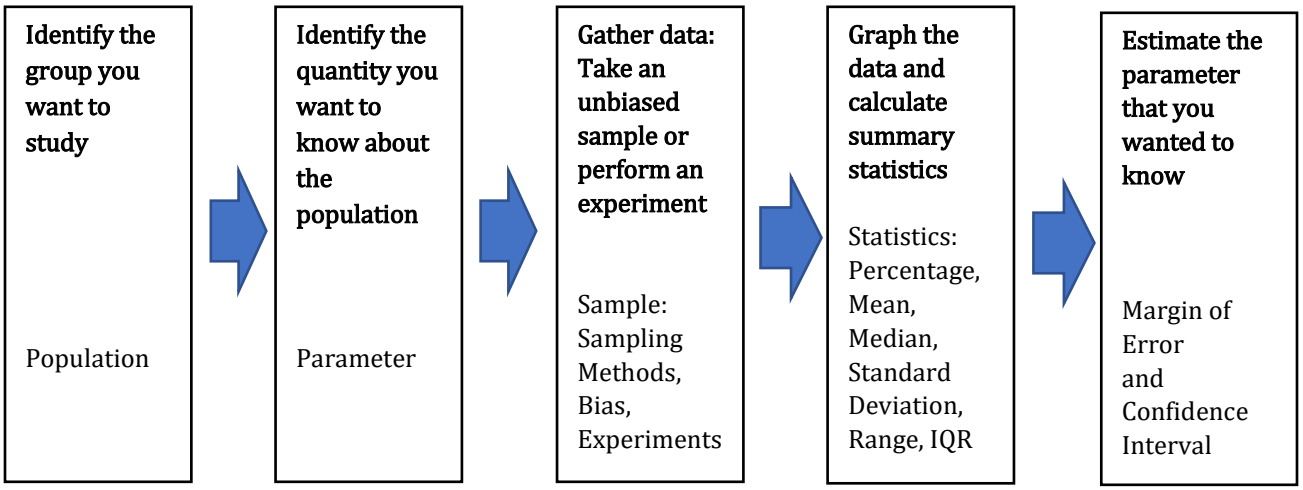
All these claims are statistical in character. We suspect that some of them sound familiar; if not, you have probably heard other claims like them. Notice how diverse the examples are; they come from psychology, health, law, sports, business, etc. Data and data-interpretation show up in virtually every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

Many of these numbers do not represent careful statistical analysis. They can be misleading and push you into decisions that you might regret. This chapter will help you learn the skills to be a critical consumer of statistical claims.

Statistical Process

To give you an overview, here is a diagram of the steps taken in a poll or other statistical study, and the elements in each step that we will discuss in this chapter. We will use many examples to illustrate the whole process.

Overview of the Statistical Process



Population

Before we begin gathering any data to analyze, we need to identify the population we are studying. The **population** of a study is the group we want to know something about. The population could be people, auto parts or tomato plants.

If we want to know the amount of money spent on textbooks by a typical college student, our population might be all students at Portland Community College. Or it might be:

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

The intended population is also called the **target population**, since if we design our study badly, the collected data might not actually be representative of the intended population.

> Example 1: A newspaper website contains a poll asking people their opinion on a recent news article. What is the population?
>
> While the target (intended) population may have been all people, the real population of the survey is readers of the website.

Parameter

A **parameter** is the value (percentage, average, etc.) that we are interested in for the whole population. Since it is often too time-consuming, expensive and/or impossible to get data for the entire population, the parameter is usually a theoretical quantity that we are trying to estimate. For example, the typical amount of money spent per year on textbooks by students at your college in a year is a parameter.

Sample

To estimate the parameter, we select a **sample**, which is a smaller subset of the entire population. It is very important that we choose a **representative sample**, one that matches the characteristics of the population, to have a good estimate. If we survey 100 students at your college, those students would be our sample. We will talk about how to choose a representative sample later in this section.

Statistic

To get our **data**, we would then ask each student in the sample how much they spent on textbooks and record the answers, or **raw data**. Then we could calculate the average, which is our statistic. A **statistic** is a value (percentage, average, etc.) calculated using data from a sample.

> Example 2: A researcher wants to know how the citizens of Portland feel about a voter initiative. To study this, they go downtown to the Pioneer Place Mall and survey 500 shoppers. Sixty percent indicate they are supportive of the initiative. Identify the population, parameter, sample and statistic.
>
> *Population:* While the intended population of this survey is Portland citizens, the effective population is Pioneer Place Mall shoppers. There is no reason to assume that shoppers at this mall would be representative of all Portland citizens.
>
> *Parameter:* The parameter is what we want to know about the population, the percentage of Portland citizens that support the initiative.
>
> *Sample:* The sample is the subgroup of the population selected. The 500 shoppers questioned make up the sample, which, again, is probably not representative of the population.
>
> *Statistic:* The statistic is calculated using the data from the sample. The percentage of people sampled who support the initiative is 60%.

Sampling

As we mentioned in a previous section, the first thing we should do before conducting a survey is to identify the population that we want to study. Suppose we are hired by a politician to determine the amount of support they have among the electorate should they decide to run for another term. What population should we study? Every person in the district? Eligible voters might be better, but what if they don't register? Registered voters may not vote. What about "likely voters?"

This is the criteria used in a lot of political polling, but it is sometimes difficult to define a "likely voter." Here is an example of the challenges of political polling.

> Example 3: In November 1998, former professional wrestler Jesse "The Body" Ventura was elected governor of Minnesota. Up until right before the election,

most polls showed he had little chance of winning. There were several contributing factors to the polls not reflecting the actual intent of the electorate:

- Ventura was running on a third-party ticket and most polling methods are better suited to a two-candidate race.
- Many respondents to polls may have been embarrassed to tell pollsters that they were planning to vote for a professional wrestler.
- The mere fact that the polls showed Ventura had little chance of winning might have prompted some people to vote for him in protest to send a message to the major-party candidates.

But one of the major contributing factors was that Ventura recruited a substantial amount of support from young people, particularly college students, who had never voted before and who registered specifically to vote in that election. The polls did not deem these young people likely voters (since in most cases young people have a lower rate of voter registration and a lower turnout rate for elections) so the polling samples were subject to **sampling bias**: they omitted a portion of the electorate that was weighted in favor of the winning candidate.

So, identifying the population can be a difficult job, but once we have identified the population, how do we choose a good sample? We want our statistic to estimate the parameter we are interested in, so we need to have a representative sample. Returning to our hypothetical job as a political pollster, we would not anticipate very accurate results if we drew all of our samples from customers at a Starbucks, or the membership list of the local Elks club. How do we get a sample that resembles our population?

Sampling Methods

One way to get a representative sample is to use *randomness*. We will look at three types of sampling that use randomness and one that does not.

Simple random sample (SRS)

A **simple random sample**, abbreviated SRS, is one in which every member of the population has an equal probability of being chosen.

Example 4: If we could somehow identify all likely voters in the state, put each of their names on a piece of paper, toss the slips into a (very large) hat and draw 1000 slips out of the hat, we would have a simple random sample.

In practice, computers are better suited for this sort of endeavor. It is always possible, however, that even a random sample might end up not being totally representative of the population. If we repeatedly take samples of 1000 people from among the population of likely voters in the state of Oregon, some of these samples might tend to have a slightly higher percentage of Democrats (or Republicans) than does the general population; some samples might include more older people and some samples might include more younger people; etc. This is called **sampling variation**.

If there are certain groups that we want to make sure are represented, we might instead use a stratified sample.

Stratified sampling

In **stratified sampling**, a population is divided into a number of subgroups (or strata). Random samples are then taken from each subgroup. It is often desirable to make the sample sizes proportional to the size of each subgroup in the population.

> Example 5: Suppose that data from voter registrations in the state indicated that the electorate was comprised of 39% Democrats, 37% Republicans and 24% Independents. In a sample of 1000 people, they would then expect to get about 390 Democrats, 370 Republicans and 240 Independents. To accomplish this, they could 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.

A way to remember stratified sampling is think about having a piece of layer cake. Each layer represents a stratum or subgroup, and a slice of the cake represents a sample of each layer.

Systematic sampling

In **systematic sampling**, every n^{th} member of the population is selected to be in the sample. The starting position is often chosen at random.

> Example 6: To select a systematic sample, Portland Community College could use their database to select a random student from the first 100 student ID numbers. Then they would select every 100^{th} student ID number after that.

Systematic sampling is not as random as a simple random sample (if your ID number is right next to your friend's because you applied at the same time, you could not both end up in the same sample) but it can yield acceptable samples. This method can be useful for people waiting in lines, parts on a manufacturing line, or plants in a row.

Convenience sampling

Convenience sampling is when samples are chosen by selecting whomever is convenient. This is the worst type of sampling because it does not use randomness.

> Example 7: A pollster stands on a street corner and interviews the first 100 people who agree to speak to them. This is a convenience sample.

Statistical Bias

There is no way to correct for biased data, so it is very important to think through the entire study and data analysis before we start. We talked about **sampling** or **selection bias**, which is when the sample is not representative of the population. One example of this is **voluntary response bias**, which is bias introduced by only collecting data from those who volunteer to participate. This can lead to bias if the people who volunteer

have different characteristics than the general population. Here is a summary of some additional sources of bias.

Types of bias

Sampling bias – when the sample is not representative of the population

> **Voluntary response bias** – the sampling bias that often occurs when the sample is made up of volunteers

Self-interest study – bias that can occur when the researchers have an interest in the outcome

Response bias – when the responder gives inaccurate responses for any reason

> **Perceived lack of anonymity** – when the responder fears giving an honest answer might negatively affect them
>
> **Loaded questions** – when the question wording influences the responses

Non-response bias – when people refuse to participate in a study or drop out of an experiment, we can no longer be certain that our sample is representative of the population

Sources of bias may be conscious or unconscious. They may be innocent or as intentional as pressuring by a pollster. Here are some examples of the types of bias.

> Example 8:
>
> a. Consider a recent study which found that chewing gum may raise math grades in teenagers[1]. This study was conducted by the Wrigley Science Institute, a branch of the Wrigley chewing gum company. This is an example of a self-interest study; one in which the researches have a vested interest in the outcome of the study. While this does not necessarily mean the study was biased, we should subject the study to extra scrutiny.
>
> b. Consider online reviews of products and businesses. Customers tend to leave reviews if they are very satisfied or very dissatisfied. While you can look for overall patterns and get useful information, these reviews suffer from voluntary response bias and likely capture more extreme views than the general population.
>
> c. A survey asks participants a question about their interactions with people of different ethnicities. This study could suffer from response bias. A respondent might give an untruthful answer to not be perceived as racist.

[1] Reuters. http://news.yahoo.com/s/nm/20090423/od_uk_nm/oukoe_uk_gum_learning. Retrieved 4/27/09

d. An employer puts out a survey asking their employees if they have a drug abuse problem and need treatment help. Here, answering truthfully might have serious consequences; responses might not be accurate if there is a perceived lack of anonymity and employees fear retribution.

e. A survey asks, "Do you support funding research on alternative energy sources to reduce our reliance on high-polluting fossil fuels?" This is an example of a loaded or leading question – questions whose wording leads the respondent towards a certain answer.

f. A poll was conducted by phone with the question, "Do you often have time to relax and read a book?" Fifty percent of the people who were called refused to participate in the survey (Probably because they didn't have the time). It is unlikely that the results will be representative of the entire population. This is an example of non-response bias.

Loaded questions can occur intentionally by pollsters with an agenda, or accidentally through poor question wording. Also of concern is question order, where the order of questions changes the results. Here is an example from a psychology researcher[2]:

> Example 9: *"My favorite finding is this: we did a study where we asked students, 'How satisfied are you with your life? How often do you have a date?' The two answers were not statistically related - you would conclude that there is no relationship between dating frequency and life satisfaction. But when we reversed the order and asked, 'How often do you have a date? How satisfied are you with your life?' the statistical relationship was a strong one. You would now conclude that there is nothing as important in a student's life as dating frequency."*

Observational Studies

So far, we have primarily discussed surveys and polls, which are types of **observational studies** – studies based on observations or measurements. These observations may be solicited, like in a survey or poll. Or, they may be unsolicited, such as studying the percentage of cars that turn right at a red light even when there is a "no turn on red" sign.

Experiments

In contrast, it is common to use **experiments** when exploring how subjects react to an outside influence. In an experiment, some kind of **treatment** is applied to the subjects and the results are measured and recorded. When conducting experiments, it is essential to isolate the treatment being tested. Here are some examples of treatments.

[2] Swartz, Norbert. http://www.umich.edu/~newsinfo/MT/01/Fal01/mt6f01.html. Retrieved 3/31/2009

Example 10:

a. A pharmaceutical company tests a new medicine for treating Alzheimer's disease by administering the drug to 50 elderly patients with recent diagnoses. The treatment here is the new drug.

b. A gym tests out a new weight loss program by enlisting 30 volunteers to try out the program. The treatment here is the new program.

c. A psychology researcher explores the effect of music on affect by measuring people's mood while listening to different types of music. The music is the treatment.

d. Suppose a middle school finds that their students are not scoring well on the state's standardized math test. They decide to run an experiment to see if a new curriculum would improve scores. To run the test, they hire a math specialist to come in and teach a class using the new curriculum. To their delight, they see an improvement in test scores.

The difficulty with the last scenario is that it is not clear whether the new curriculum or the math specialist is responsible for the improvement. This is called confounding and it is the downfall of many experiments, though it is often hidden.

Confounding

Confounding occurs when there are two or more potential variables that could have caused the outcome and it is not possible to determine which one actually caused the result.

Example 11:

a. A drug company study about a weight loss pill might report that people lost an average of 8 pounds while using their new drug. However, in the fine print you find a statement saying that participants were encouraged to also diet and exercise. It is not clear in this case whether the weight loss is due to the pill, to diet and exercise, or a combination of both. In this case confounding has occurred.

b. Researchers conduct an experiment to determine whether students will perform better on an arithmetic test if they listen to music during the test. They first give the student a test without music, then give a similar test while the student listens to music. In this case, the student might perform better on the second test, regardless of the music, simply because it was the second test and they were warmed up.

There are a number of measures that can be introduced to help reduce the likelihood of confounding. The primary measure is to use a control group.

Control group

In experiments, the participants are typically divided into a treatment group and a control group. The treatment group receives the treatment being tested; the **control group** does not receive the treatment.

Ideally, the groups are otherwise as similar as possible, isolating the treatment as the only potential source of difference between the groups. For this reason, the method of dividing groups is important. Some researchers attempt to ensure that the groups have similar characteristics (same number of each gender identity, same number of people over 50, etc.), but it is nearly impossible to control for every characteristic. Because of this, random assignment is very commonly used.

Example 12:

a. To determine if a two-day prep course would help high school students improve their scores on the SAT test, a group of students was randomly divided into two subgroups. The first group, the treatment group, was given a two-day prep course. The second group, the control group, was not given the prep course. Afterwards, both groups took the SAT test.

b. A company testing a new plant food grows two crops of plants in adjacent fields that typically produce the same amount of food. The treatment group receives the new plant food and the control group does not. The crop yields would then be compared. By growing the two crops at the same time in similar fields, they are controlling for weather and other confounding factors.

Sometimes not giving the control group anything does not completely control for confounding variables. For example, suppose a medicine study is testing a new headache pill by giving the treatment group the pill and the control group nothing. If the treatment group showed improvement, we would not know whether it was due to the medicine, or a response to have something. This is called a placebo effect.

Placebo effect

The **placebo effect** is when the effectiveness of a treatment is influenced by the patient's perception of how effective they think the treatment will be, so a result might be seen even if the treatment is ineffectual.

Example 13: A study found that when doing painful dental tooth extractions, patients told they were receiving a strong painkiller while actually receiving a saltwater injection found as much pain relief as patients receiving a dose of morphine.[3]

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

To control for the placebo effect, a **placebo**, or dummy treatment, is often given to the control group. This way, both groups are truly identical except for the specific treatment given.

Placebo and Placebo-controlled experiments

An experiment that gives the control group a placebo is called a **placebo-controlled** experiment.

Example 14:

a. In a study for a new medicine that is dispensed in a pill form, a sugar pill could be used as a placebo.

b. In a study on the effect of alcohol on memory, a non-alcoholic beer might be given to the control group as a placebo.

c. In a study of a frozen meal diet plan, the treatment group would receive the diet food, and the control group could be given standard frozen meals taken out of their original packaging.

In some cases, it is more appropriate to compare to a conventional treatment than a placebo. For example, in a cancer research study, it would not be ethical to deny any treatment to the control group or to give a placebo treatment. In this case, the currently acceptable medicine would be given to the second group, called a **comparison group**. In our SAT test example, the non-treatment group would most likely be encouraged to study on their own, rather than be asked to not study at all, to provide a meaningful comparison. It is very important to consider the ethical ramifications of any experiment.

Blind studies

A **blind study** is one that uses a placebo and the participants do not know whether they are receiving the treatment or a placebo. A **double-blind study** is one in which the subjects and those interacting with them don't know who is in the treatment group and who is in the control group.

Example 15: In a study about anti-depression medication, you would not want the psychological evaluator to know whether a patient is in the treatment or control group, as it might influence their evaluation. The experiment should be conducted as a double-blind study.

Margin of Error and Confidence Intervals

Even when a study or experiment has successfully avoided bias and has been well done, there is still an element of variation. If we took 5 different random samples of 100 college students and calculated their average textbook cost, we wouldn't expect to get the exact same average for each sample. This is due to sampling variation. To account for this, researchers publish their margin of error or a confidence interval for their

statistics. These numbers describe the precision of the estimate for a certain confidence level.

You've probably heard something like, "The candidate has 54 percent of the likely voters, plus or minus three percent." The 3% is called the **margin of error**, so the true percentage is somewhere between 51% and 57%, with a certain level of confidence. To write this as a **confidence interval**, we place the numbers in parentheses from smallest to largest, separated by a comma: (51%, 57%).

The most common **confidence level** is 95%, which means if the poll was conducted repeatedly, and we made a confidence interval each time, we would expect the true percentage, or parameter, to fall within our confidence interval 95 out of 100 times. You can learn more on how to calculate the margin of error for different confidence levels in a statistics class.

> Example 16: Let's say we asked a random sample of 100 students at Portland Community College and found that they spent an average of $451.32 on books their first year, plus or minus $85.63. Write this as a confidence interval, assuming a 95% confidence level.
>
> If the margin of error was calculated to be plus or minus $85.63, then with a confidence level of 95% we could say that the average amount spent by the population is somewhere between $424.69 and $478.26. We could also write this as a **confidence interval**: ($365.69, $536.95).

Now we have come full circle and seen how we can use data from a sample to estimate the parameter we were interested in for our population.

Exercises 3.1

1. A political scientist surveys 28 of the current 106 representatives in a state's congress. Of them, 14 said they were supporting a new education bill, 12 said there were not supporting the bill, and 2 were undecided.
    a. Who is the population of this survey?
    b. What is the size of the population?
    c. What is the size of the sample?
    d. Give the statistic for the percentage of representatives surveyed who said they were supporting the education bill.
    e. If the margin of error was 5%, give the confidence interval for the percentage of representatives we might we expect to support the education bill.

2. The city of Raleigh has 9,500 registered voters. There are two candidates for city council in an upcoming election: Brown and Feliz. The day before the election, a telephone poll of 350 randomly selected registered voters was conducted. 112 said they'd vote for Brown, 207 said they'd vote for Feliz, and 31 were undecided.
    a. Who is the population of this survey?

   b. What is the size of the population?
   c. What is the size of the sample?
   d. Give the statistic for the percentage of voters surveyed who said they'd vote for Brown.
   e. If the margin of error was 3.5%, give the confidence interval for the percentage of voters surveyed that we might we expect to vote for Brown.

3. To determine the average length of trout in a lake, researchers catch 20 fish and measure them. Describe the population and sample of this study.

4. A college reports that the average age of their students is 28 years old. Is this a parameter or a statistic?

5. Which sampling method is being described?
   a. In a study, the sample is chosen by separating all cars by size and selecting 10 of each size grouping.
   b. In a study, the sample is chosen by writing everyone's name on a playing card, shuffling the deck, then choosing the top 20 cards.
   c. Every 4th person on the class roster was selected.

6. Which sampling method is being described?
   a. A sample was selected to contain 25 people aged 18-34 and 30 people aged 35-70.
   b. Viewers of a new show are asked to respond to a poll on the show's website.
   c. To survey voters in a town, a polling company randomly selects 100 addresses from a database and interviews those residents.

7. Identify the most relevant source of bias in each situation.
   a. A survey asks the following: Should the mall prohibit loud and annoying rock music in clothing stores catering to teenagers?
   b. To determine opinions on voter support for a downtown renovation project, a surveyor randomly questions people working in downtown businesses.
   c. A survey asks people to report their actual income and the income they reported on their IRS tax form.
   d. A survey randomly calls people from the phone book and asks them to answer a long series of questions.
   e. The Beef Council releases a study stating that consuming red meat poses little cardiovascular risk.
   f. A poll asks, "Do you support a new transportation tax, or would you prefer to see our public transportation system fall apart?"

8. Identify the most relevant source of bias in each situation.
    a. A survey asks the following: Should the death penalty be permitted if innocent people might die?
    b. A study seeks to investigate whether a new pain medication is safe to market to the public. They test by randomly selecting 300 people who identify as men from a set of volunteers.
    c. A survey asks how many sexual partners a person has had in the last year.
    d. A radio station asks listeners to phone in their response to a daily poll.
    e. A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score.
    f. High school students are asked if they have consumed alcohol in the last two weeks.

9. Identify whether each situation describes an observational study or an experiment.
    a. The temperature on randomly selected days throughout the year was measured.
    b. One group of students listened to music and another group did not while they took a test and their scores were recorded.
    c. The weights of 30 randomly selected people are measured.

10. Identify whether each situation describes an observational study or an experiment.
    a. Subjects are asked to do 20 jumping jacks, and then their heart rates are measured.
    b. Twenty coffee drinkers and twenty tea drinkers are given a concentration test.
    c. The weights of potato chip bags are weighed on the production line before they are put into boxes.

11. A team of researchers is testing the effectiveness of a new vaccine for human papilloma virus (HPV). They randomly divide the subjects into two groups. Group 1 receives new HPV vaccine, and Group 2 receives the existing HPV vaccine. The patients in the study do not know which group they are in.
    a. Which is the treatment group?
    b. Which is the control group (if there is one)?
    c. Is this study blind, double-blind, or neither?
    d. Is this best described as an experiment, a controlled experiment, or a placebo-controlled experiment?

12. For the clinical trials of a weight loss drug containing *Garcinia Cambogia* the subjects were randomly divided into two groups. The first received an inert pill along with an exercise and diet plan, while the second received the test medicine along with the same exercise and diet plan. The patients do not know which group they are in, nor do the fitness and nutrition advisors.
    a. Which is the treatment group?
    b. Which is the control group (if there is one)?
    c. Is this study blind, double-blind, or neither?
    d. Is this best described as an experiment, a controlled experiment, or a placebo-controlled experiment?

13. A study is conducted to determine whether people learn better with routine or crammed studying. Subjects volunteer from an introductory psychology class. At the beginning of the semester 12 subjects volunteer and are assigned to the routine studying group. At the end of the semester 12 subjects volunteer and are assigned to the crammed studying group.
    a. Identify the target population and the sample.
    b. Is this an observational study or an experiment?
    c. This study involves two kinds of non-random sampling: 1. Subjects are not randomly sampled from a specified population and 2. Subjects are not randomly assigned to groups. Which problem is more serious? What effect on the results does each have?

14. A farmer believes that playing Barry Manilow songs to his peas will increase their yield. Describe a controlled experiment the farmer could use to test his theory.

15. A sports psychologist believes that people are more likely to be extroverted as an adult if they played team sports as a child. Describe two possible studies to test this theory. Design one as an observational study and the other as an experiment. Which is more practical?

16. To test a new lie detector, two groups of subjects are given the new test. One group is asked to answer all the questions truthfully. The second group is asked to tell the truth on the first half of the questions and lie on the second half. The person administering the lie detector test does not know what group each subject is in. Does this experiment have a control group? Is it blind, double-blind, or neither? Explain.

17. A poll found that 30%, plus or minus 5% of college freshmen prefer morning classes to afternoon classes.
    a. What is the margin of error?
    b. Write the survey results as a confidence interval.

18. A poll found that 38% of U.S. employees are engaged at work, plus or minus 3.5%.
    a. What is the margin of error?
    b. Write the survey results as a confidence interval.

19. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer treatment is under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout this new treatment.
    a. What is the population of this study?
    b. Would you expect the data from the two researchers to be identical? Why or why not?
    c. If the first researcher collected their data by randomly selecting 10 nearby ZIP codes, then selecting 4 people from each, which sampling method did they use?
    d. If the second researcher collected their data by choosing 40 patients they knew, what sampling method did they use? What concerns would you have about this data set, based upon the data collection method?

20. Find a newspaper or magazine article, or the online equivalent, describing the results of a recent study (not a simple poll). Give a summary of the study's findings, then analyze whether the article provided enough information to determine the validity of the conclusions. If not, produce a list of things that are missing from the article that would help you determine the validity of the study. Look for the things discussed in the text: population, sample, randomness, blind, control, margin of error, etc.

21. Use a polling website such as www.pewresearch.org or www.gallup.com and search for a poll that interests you. Find the result, the margin of error and confidence level for the poll and write the confidence interval.

Section 3.2 Describing Data

Once we have collected data from an observational study or an experiment, we need to summarize and present it in a way that will be meaningful to our audience. The raw data is not very useful by itself. In this section we will begin with graphical presentations of data and in the rest of the chapter we will learn about numerical summaries of data.

Types of Data

There are two types of data, categorical data and quantitative data.

Categorical (qualitative) data are pieces of information that allow us to classify the subjects into various categories.

> Example 1: We might conduct a survey to determine the name of the favorite movie that people saw in a movie theater. When we conduct such a survey, the responses would look like: *Finding Nemo*, *Black Panther*, *Titanic, etc*.
>
> We can count the number of people who give each answer, but the answers themselves do not have any numerical values: we cannot perform computations with an answer like *"Black Panther"* because it is categorical data.

Quantitative data are responses that are numerical in nature and with which we can perform meaningful calculations.

> Example 2: A survey could ask the number of movies you have seen in a movie theater in the past 12 months (0, 1, 2, 3, 4, ...). This would be quantitative data.

Other examples of quantitative data would be the running time of the movie you saw most recently (104 minutes, 137 minutes, 110 minutes, etc.) or the amount of money you paid for a movie ticket the last time you went to a movie theater ($5.50, $9.75, $10.50, etc.).

We cannot assume that all numbers are quantitative data, and sometimes it is not so clear-cut. Here are some examples to illustrate this.

> Example 3:
>
> a. Suppose we gather respondents' ZIP codes in a survey to track their geographical location. ZIP codes are numbers, but we can't do any meaningful calculations with them (it doesn't make sense to say that 98036 is "twice" 49018 — that's like saying that Lynnwood, WA is "twice" Battle Creek, MI, which doesn't make sense at all), so ZIP codes are really categorical data.
>
> b. A survey about the movie you most recently saw includes the question, "How would you rate the movie?" with these possible answers:

1 - It was awful.
2 - It was just okay.
3 - I liked it.
4 - It was great.
5 - Best movie ever!

Again, there are numbers associated with the responses, but these are really categories. A movie that rates a 4 is not necessarily twice as good as a movie that rates a 2, whatever that means; However, we often see that a movie got an average of 3.7 stars, which is an average of categorical ratings and it can give us important information.

Overall, it is important to look at the purpose of the study for any variables that could be classified as either categorical or quantitative. Another consideration is what you plan to do with the data. Next, we will talk about how to display each type of data.

Presenting Categorical Data

Since we can't do calculations with categorical data, we begin by summarizing the data in a frequency table or a relative frequency table.

Frequency Tables

A **frequency table** has one column for the categories, and another for the **frequency**, or number of times that category occurred.

Example 4: An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get in accidents. To research this, they examine police reports for recent total-loss collisions. The data is summarized in this table.

Car Color	Frequency of Total-Loss Collisions
Blue	25
Green	52
Red	41
White	36
Black	39
Grey	23
Total	**216**

Relative Frequency Tables

Numbers are usually not as easy to interpret as percentages, so we will add a column for the relative frequencies. A **relative frequency** is the percentage for the category, found by dividing each frequency by the total and converting to a percentage. You'll notice the percentages may not add up to exactly 100% due to rounding.

Example 4 Continued:

Car Color	Frequency of Total-Loss Collisions	Relative Frequency of Total-Loss Collisions
Blue	25	25/216 = 0.116 or 11.6%
Green	52	52/216 = 0.241 or 24.1%
Red	41	41/216 = 0.190 or 19.0%
White	36	36/216 = 0.167 or 16.7%
Black	39	39/216 = 0.181 or 18.1%
Grey	23	23/216 = 0.107 or 10.7%
Total	**216**	216/216 = 1.0 or 100%

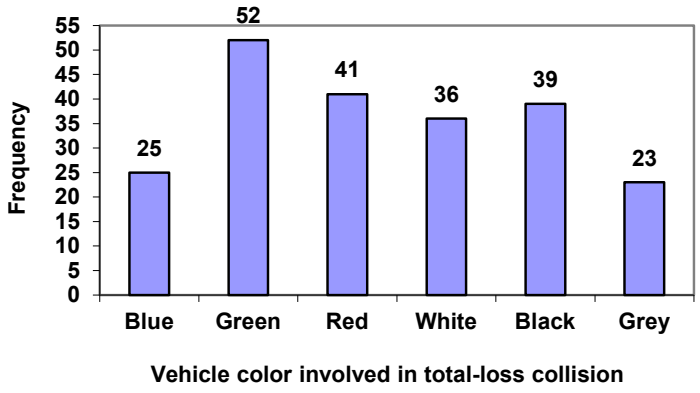
It would be even more useful to have a visual to see what is going on, and this is where charts and graphs come in. For categorical data we can display our data using bar graphs and pie charts.

Bar graphs

A **bar graph** is a graph that displays a bar for each category with the height of the bar indicating the frequency of that category. To construct a bar graph with vertical bars, we label the horizontal axis with the categories. The vertical axis will have a scale for the frequency or relative frequency.

The highest frequency in our car data is 52 collisions, so we will set our vertical axis to go from 0 to 55, with a scale of 5 units.

To draw bar graphs by hand graph paper is useful, or you can use technology. It is also very helpful to label each bar with the frequency or relative frequency.

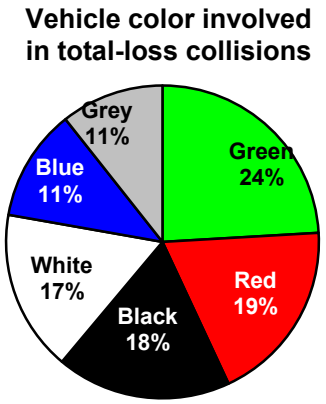


Pie Charts

A natural way to visualize relative frequencies is with a pie chart. A **pie chart** is a circle with wedges cut of varying sizes like slices of pizza or pie. The size of each wedge corresponds to the relative frequency of the category. The slices add up to 100%, just like relative frequencies.

Pie charts can often benefit from including frequencies or relative frequencies in the pie slices.

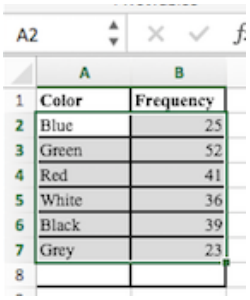
Pie charts look nice but are harder to draw by hand than bar charts since to draw them accurately we would need to compute the angle each wedge cuts out of the circle, then measure the angle with a protractor. A spreadsheet is much better suited to drawing pie charts.



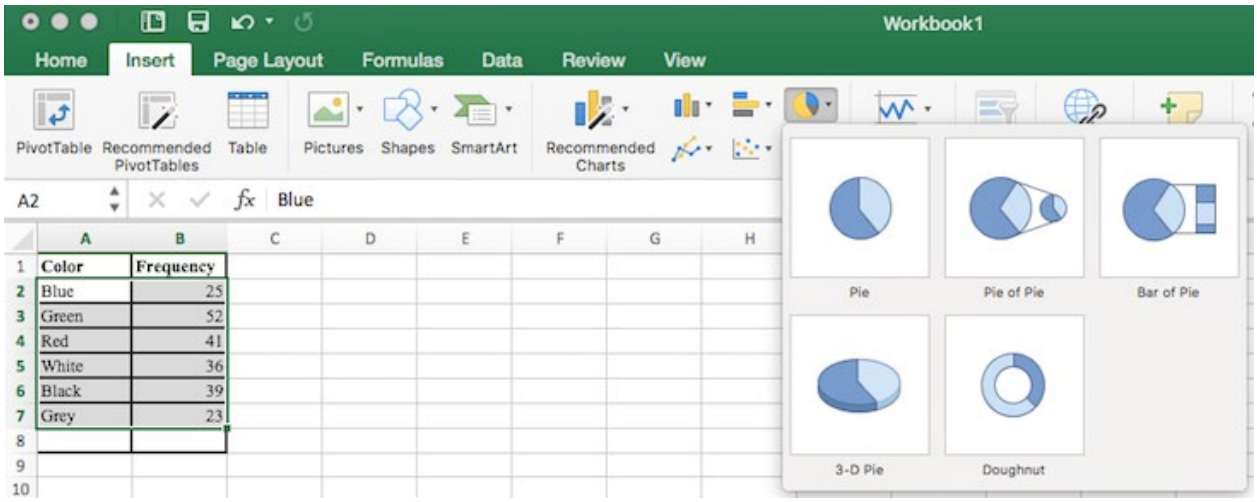
Using a Spreadsheet to Make Bar Charts and Pie Charts

To make a graph using a spreadsheet, place the data from the frequency table into the cells.

Then select the data, go to the Insert tab, and choose the bar graph or pie chart that you would like. For this example, we will choose a pie graph.

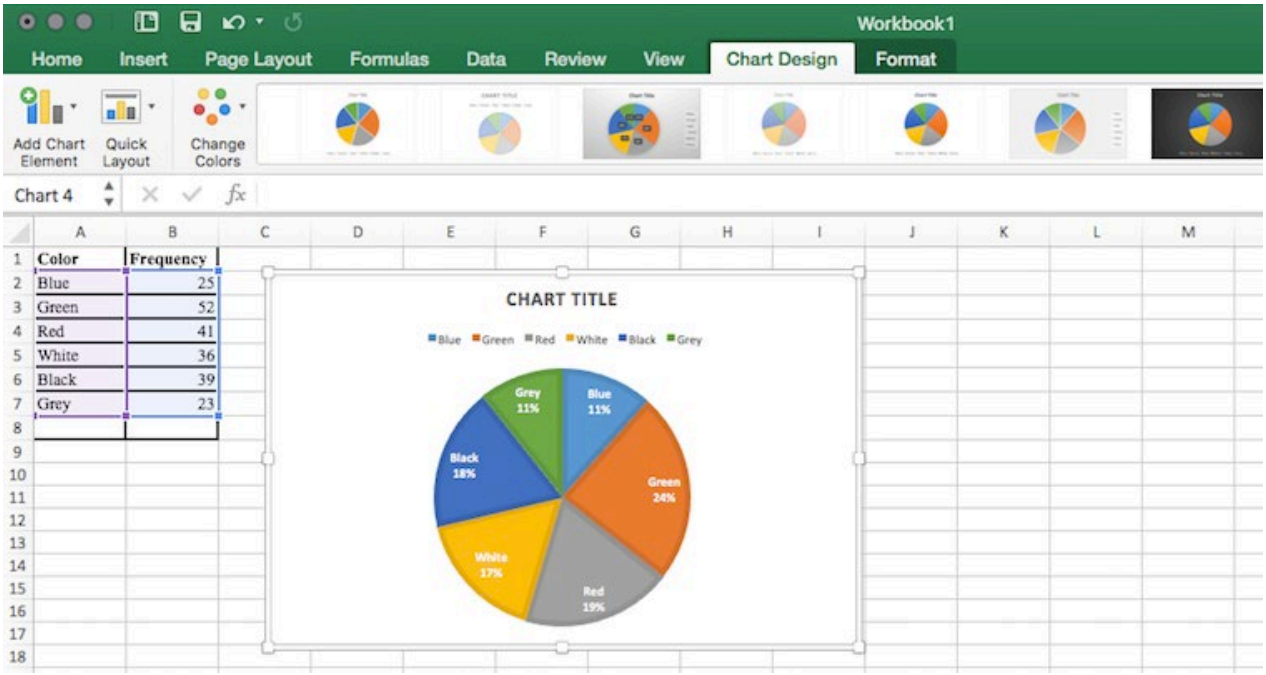


	A	B
1	Color	Frequency
2	Blue	25
3	Green	52
4	Red	41
5	White	36
6	Black	39
7	Grey	23
8		



After the spreadsheet has created your pie graph you can choose which design you prefer by clicking on the Chart Design tab. Since these pie pieces represent car colors, we matched the color of each wedge to the color of the car in our pie chart above.

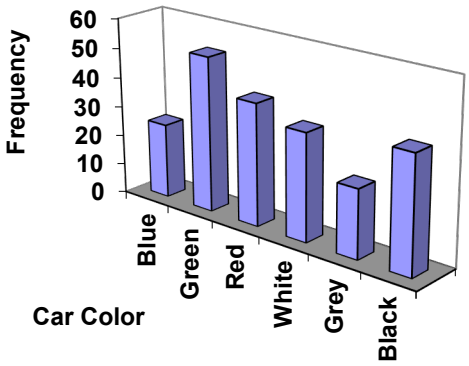
To give your graph a meaningful title, click on Chart Title. There are many other settings that you can experiment with.



Misleading Graphs

Graphs can be misleading intentionally or unintentionally. It's better to keep them simple, clear and well-labeled. People sometimes add features to graphs that don't help convey their information.

Example 5: A 3-dimensional bar chart like the one shown is usually not as effective as a 2-dimensional graph. The extra dimension does not add any useful information.



.

Here is another way that fanciness can sometimes lead to trouble. Instead of plain bars, it is tempting to substitute images. This type of graph is called a pictogram.

Perceptual Distortion

A **pictogram** is a statistical graphic in which the size of the picture is intended to represent the frequency or size of the values being represented. We need to be careful

with these, because our brains perceive the relationship between the areas, not the heights.

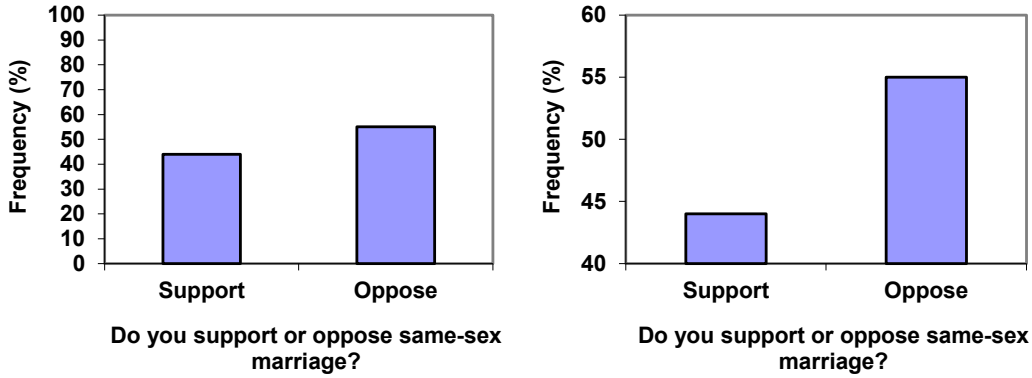
Example 6: A labor union might produce this graph to show the difference between the average manager salary and the average worker salary.

The average manager salary is twice as high as the average worker salary as in a bar graph, but the image is also twice as wide. That makes it look like the manager salary is 4 times as large as the worker salary. The area needs to accurately portray the relationship, otherwise we will have a perceptual distortion.

Misleading Scale

Another type of distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the vertical axis, representing the least number of cases that could have occurred in a category. Normally, this number should be zero.

Example 7: Compare the two graphs below showing support for same-sex marriage rights from a poll taken in December, 2008[4]. At a glance, the two graphs suggest very different stories. The second graph makes it look like more than three times as many people oppose marriage rights as support them. But when we look at the scale we can see that the difference is about 12%. By not starting at zero the difference looks enlarged.



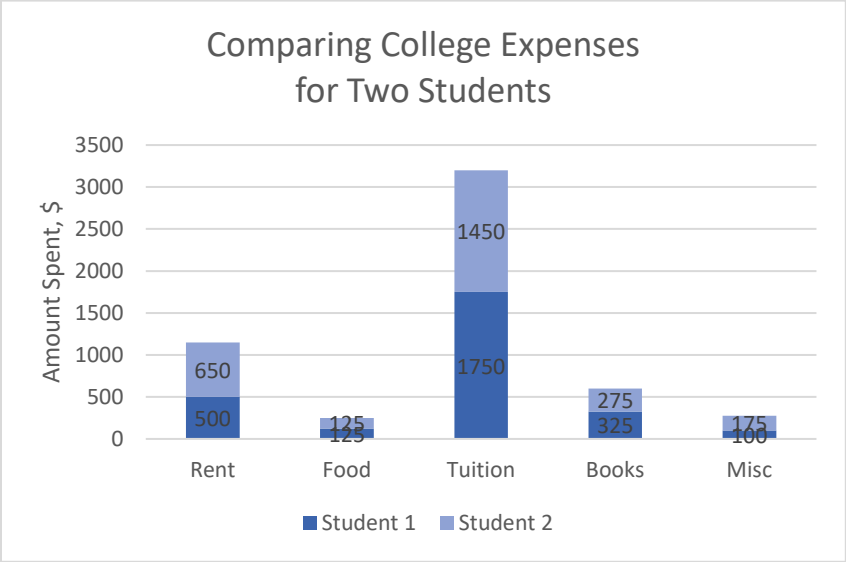
[4]CNN/Opinion Research Corporation Poll. Dec 19-21, 2008, from http://www.pollingreport.com/civil.htm

Stacked Bar Graphs

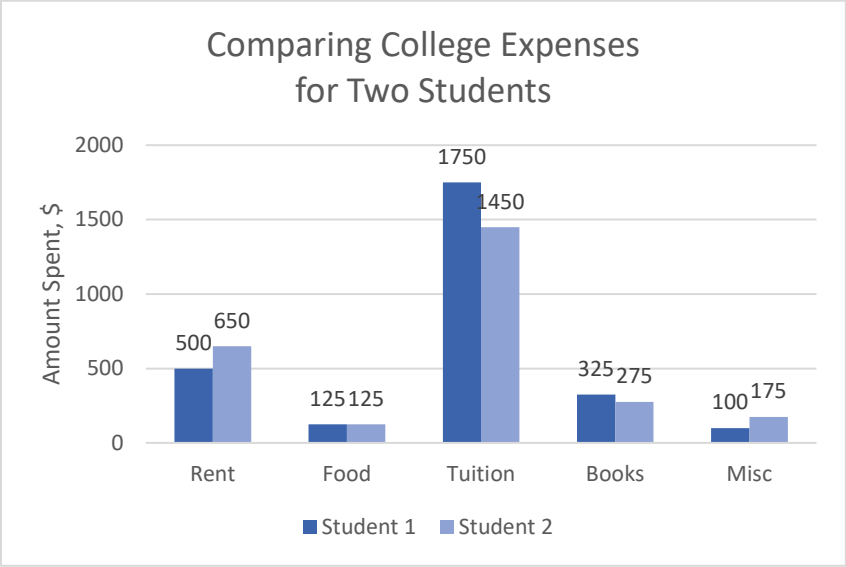
Another type of graph that can be hard to read and sometimes misleading is a stacked bar graph. In a **stacked bar graph**, the values we are comparing are stacked on top of each other vertically.

Example 8: The table lists college expenses for two different students and we want to compare them. A stacked bar graph shows the expenses stacked vertically, but we are interested in the differences, not the totals.

Expense	Student 1	Student 2
Rent	$500	$650
Food	$125	$125
Tuition	$1750	$1450
Books	$325	$275
Misc	$100	$175



It would be much easier to interpret the differences in a side-by-side bar chart.



Presenting Quantitative Data

With categorical data, the horizontal axis is the category, but with quantitative, or numerical, data we have numbers. If we have repeated values we can also make a frequency table.

Example 9: A teacher records scores on a 20-point quiz for the 30 students in their class. The scores are:

19, 20, 18, 18, 17, 18, 19, 17, 20, 18, 20, 16, 20, 15, 17, 12, 18, 19, 18, 19, 17, 20, 18, 16, 15, 18, 20, 5, 0 and 0.

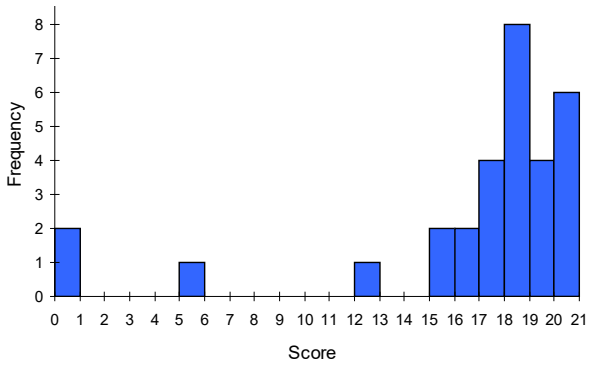
Here is a frequency table with the scores grouped and put in order.

Quiz Score	Frequency of Students
0	2
5	1
12	1
15	2
16	2
17	4
18	8
19	4
20	6

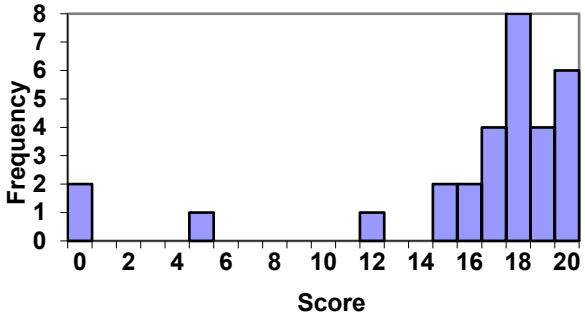
Using this table, it would be possible to create a standard bar chart from this summary, like we did for categorical data. However, since the scores are numerical values, this chart wouldn't make sense; the first and second bars would be five values apart, while the later bars would only be one value apart. Instead, we will to treat the horizontal axis as a number line. This type of graph is called a histogram.

Histograms

A **histogram** is like a bar graph, but the horizontal axis is a number line. Unlike a bar graph, there are no spaces between the bars. Here is a histogram for the data given above. Notice that in this histogram, the two scores of 15 are to the right of 15, or between 15 and 16.



The horizontal scales on histograms can be confusing for this reason. Some people choose to have bars start at ½ values to avoid this ambiguity, as in this second histogram.



If we have a large number of different data values, a frequency table listing every possible value would be way too long. There would be too many bars on the histogram to reveal any patterns. For this reason, it is common with quantitative data to group data into class intervals.

Class Intervals

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

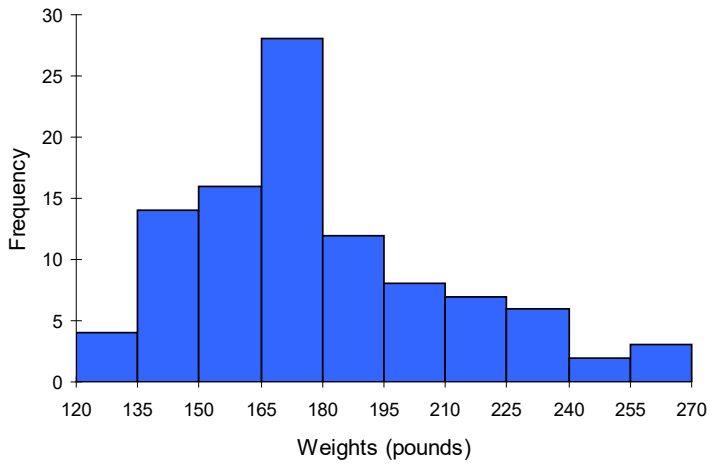
1. Each interval is equal in size. For example, if the first class contains values from 120-134, the second class should include values from 135-149.
2. We typically have somewhere between 5 and 20 classes, depending on the number of data values we're working with.

In the next example, we'll make a histogram using class intervals.

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

Interval	Frequency
120 - 134	4
135 – 149	14
150 – 164	16
165 – 179	28
180 – 194	12
195 – 209	8
210 – 224	7
225 – 239	6
240 – 254	2
255 - 269	3

Here is a histogram of this data:

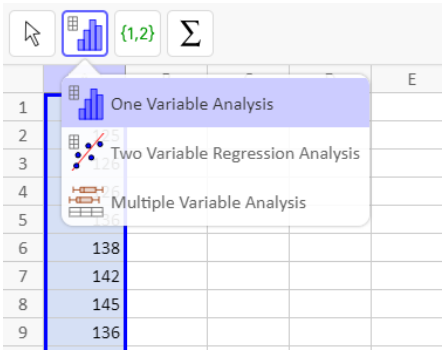


When using class intervals, it is much easier to use technology that was specifically designed to make histograms. GeoGebra is one program that lets you adjust the class widths to see which graph best displays the data.

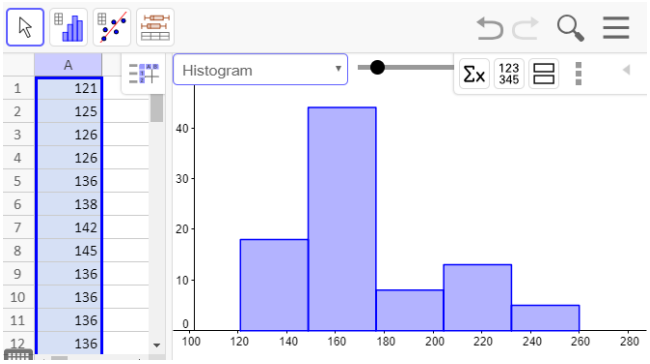
Histograms Using Technology

We will be using GeoGebra throughout this chapter to make graphs and calculate summary statistics. There is an online version and one you can download available at www.GeoGebra.org. The instructions are similar for both.

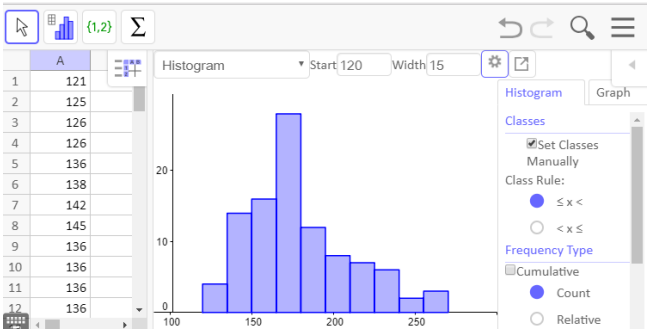
The first thing we need to do is enter the data in GeoGebra's spreadsheet. You can access the spreadsheet from Main Menu →View →Spreadsheet. Next, enter your data and select that column. Then click on the histogram icon in the menu bar on the left side and select One Variable Analysis.



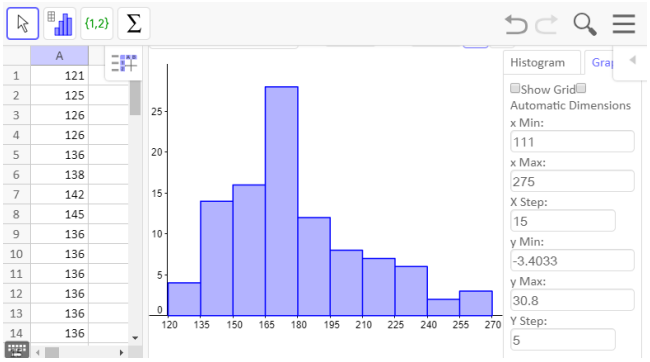
A new window will pop up showing a visual of the data. There is a drop-down menu for the type of graph, but histogram is the default. Notice that the bars are not lined up with the tick-marks at the bottom, so we want to edit this histogram. The slider bar at the top will let you see different class widths, but we want to choose our class widths manually.



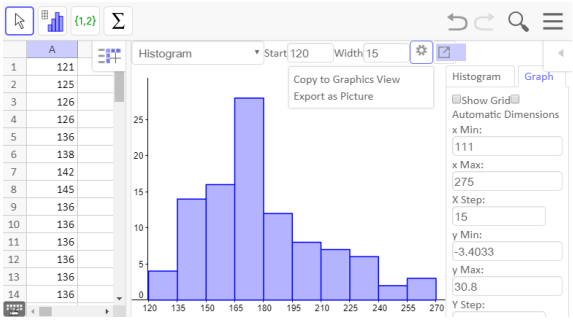
If you close the menu at the top right by clicking on the left pointing triangle, you will see a settings wheel. Click on the wheel and check the box for set classes manually. To match our previous histogram, we will start at 120 pounds and set a class width of 15 pounds.



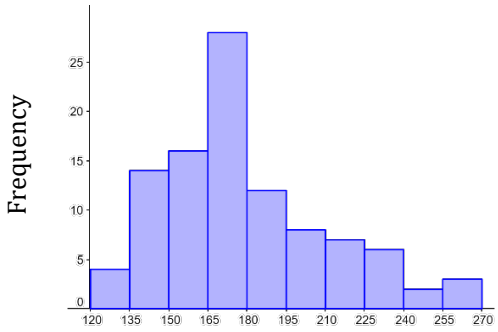
Now the bars of the histogram match our previous graph, but we need to edit the axis labels to match. Click on the graph tab on the right side and uncheck the box for automatic dimensions. We set the x min, x max, x step, y min, y max and y step as shown.



To put the graph in an assignment or a book such as this one, select the export icon and choose Export as Picture. The downloaded version also has a Copy to Clipboard option. Then insert the graph into any document and add axis labels.



Here is our finished histogram:



Weights (pounds)

The Shape of a Distribution

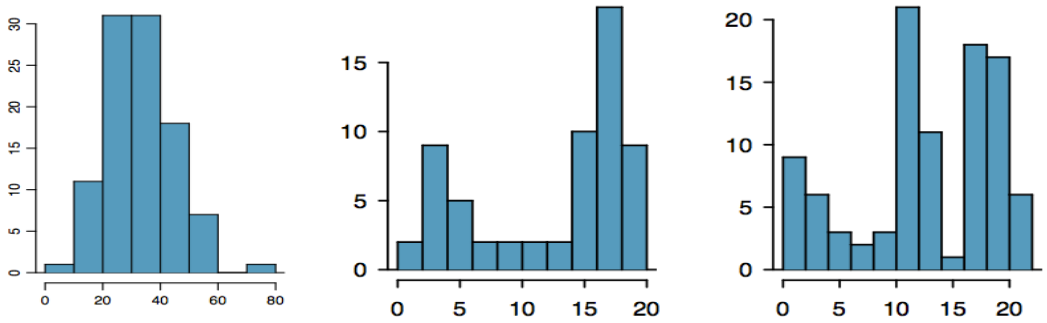
Once we have our histogram, we can use it to determine the shape of the data or **distribution**. When describing distributions, we are going to look at four characteristics: shape, center, spread and outliers. Center and spread (variation) will be covered in the next two sections.

Modality

The **modality** of a distribution indicates the number of peaks or hills in its histogram.

- It is **unimodal** if it has one peak.
- It is **bimodal** if it has two peaks.
- It is **multimodal** if it has multiple peaks.

Example 11: The first graph is unimodal, the second is bimodal and the third is multimodal.

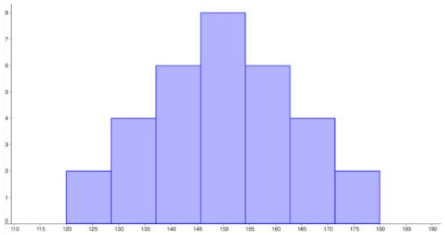


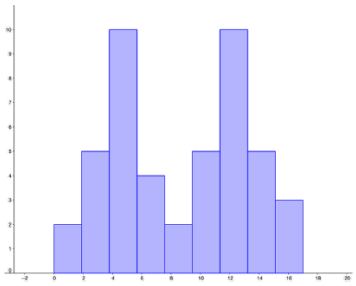
A bimodal distribution can result when two different populations have been grouped together and they are overlapping. It would be better to separate them into two separate graphs. For example, the grams of sugar per serving in sugar and non-sugar cereals.

Symmetry

A distribution is **symmetric** if the left side of the graph mirrors the right side.

> Example 12: The graph on the left is symmetric and unimodal while the graph on the right is roughly symmetric and bimodal.



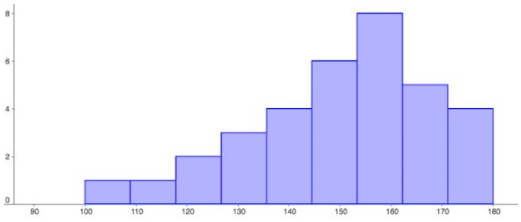


Skewness

If a distribution is not symmetric then we say it is skewed. A graph can be **skewed to the left** or **skewed to the right**. We say it is skewed in the direction of the longer tail.

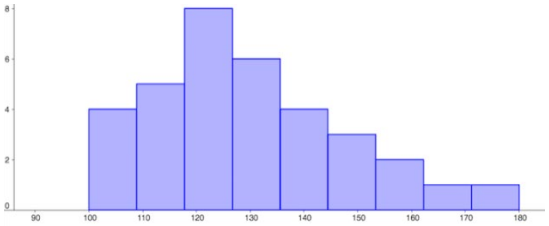
Skewed to the Left

A left skewed graph is also called a **negatively skewed** graph. The longer tail will be on the left or negative side.



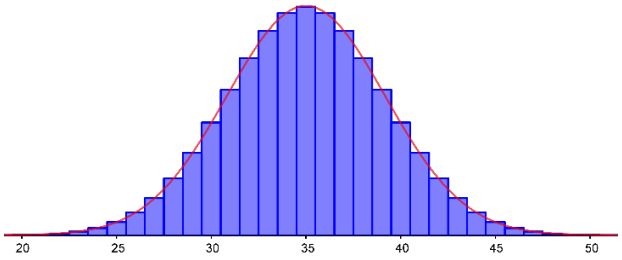
Skewed to the Right

A right skewed graph is also called a **positively skewed** graph. The longer tail will be on the right or positive side.



The Normal Distribution

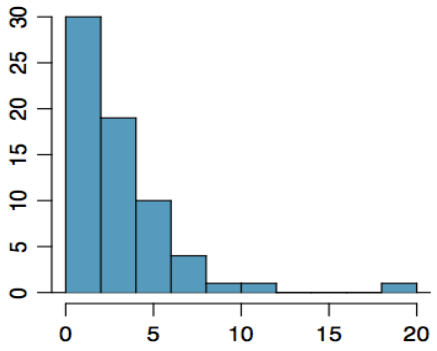
The **normal distribution** has a very specific shape. It is unimodal and symmetric with a bell-shaped graph.



Outliers

Outliers are data values that are unusually far away from the rest of the data. There is often a gap between the outlier and the rest of the graph. This visual determination of outliers is often subjective and depends on the situation.

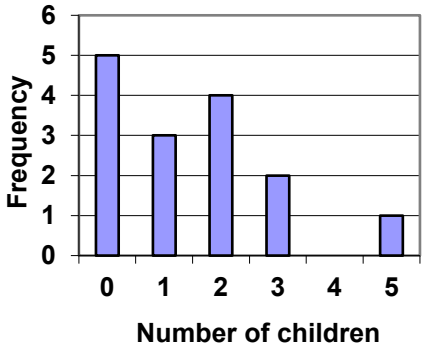
Example 13: In the graph to the right we have a unimodal distribution that is skewed to the right. There appears to be an outlier near 20.



Exercises 3.2

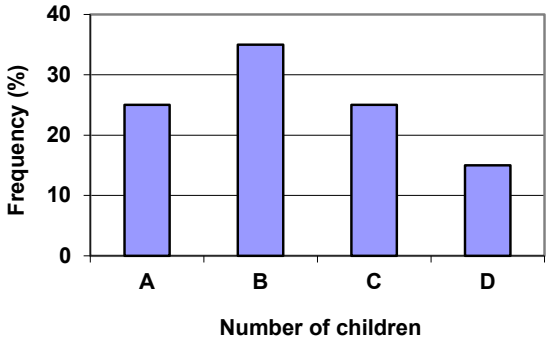
1. Is the data described categorial or quantitative?
   a. In a study, you ask the subjects their age in years.
   b. In a study, you ask the subjects their gender.
   c. In a study, you ask the subjects their ethnicity.
   d. The daily high temperature of a city over several weeks.
   e. A person's annual income.

2. A group of adults were asked how many children they have in their family. The bar graph to the right shows the number of adults who indicated each number of children.
   a. How many adults had 3 children?
   b. How many adults where questioned?
   c. What percentage of the adults questioned had 0 children?

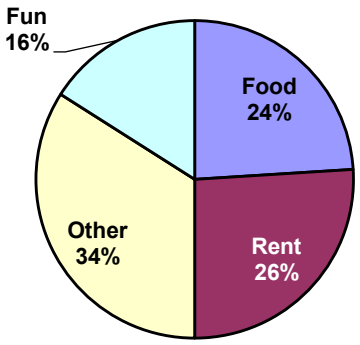


3. Jasmine was interested in how many days it would take a DVD order from Netflix to arrive at her door. The graph shows the data she collected.
   a. How many movies took 2 days to arrive?
   b. How many movies did she order in total?
   c. What percentage of the movies arrived in one day?

4. This relative frequency 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?



5. Corey categorized his spending for this month into four categories: Rent, Food, Fun, and Other. The percentages he spent in each category are pictured here. If he spent a total of $2,600 this month, how much did he spend on rent?



6. In a survey[5], 1012 adults were asked whether they personally worried about a variety of environmental concerns. The number of people who indicated that they worried "a great deal" about some selected concerns is listed below.
   a. Is this categorical or quantitative data?
   b. Make a bar chart for this data.
   c. Why can't we make a pie chart for this data?

Environmental Issue	Frequency
Pollution of drinking water	597
Contamination of soil and water by toxic waste	526
Air pollution	455
Global warming	354

7. A group of adults were asked how many cars they had in their household.
   a. Is this categorical or quantitative data?
   b. Make a relative frequency table for the data.
   c. Make a bar chart for the data.
   d. Make a pie chart for the data.

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

[5] Gallup Poll. March 5-8, 2009. http://www.pollingreport.com/enviro.htm

8. The table below shows scores on a math test.
   a. Is this categorical or quantitative data?
   b. Make a relative frequency table for the data using a class width of 10.
   c. Construct a histogram of the data.

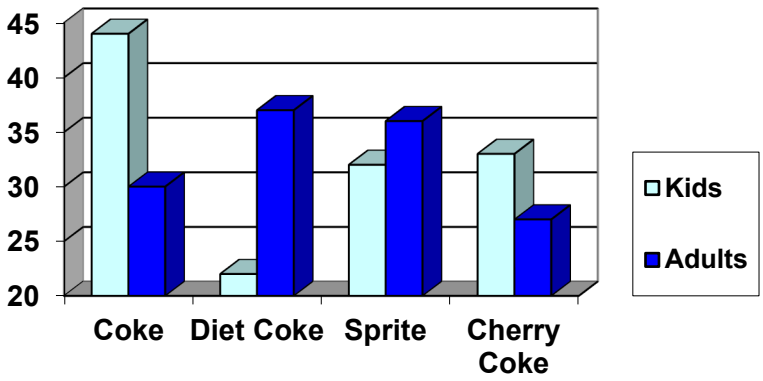
82	55	51	97	73	79	100	60	71	85	78	59
90	100	88	72	46	82	89	70	100	68	61	52

9. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer drug is currently under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout their treatment. The following data (in months) are collected.
   a. Create a histogram for each dataset, using the same class intervals and scales so you can compare them.
   b. Compare and contrast the two distributions.

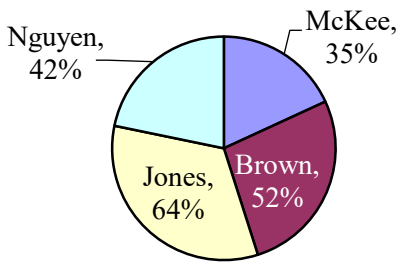
   Researcher 1: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

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

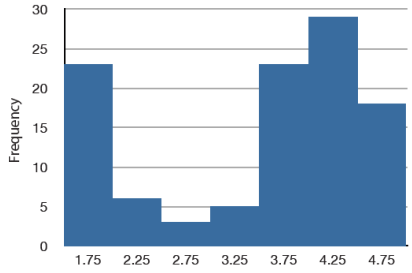
10. This graph shows the number of adults and kids who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph could be improved.



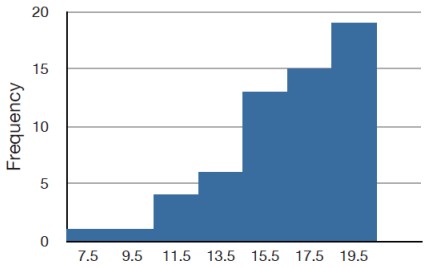
11. A poll was taken asking people if they agreed with the positions of the 4 candidates for a county office. Does this pie chart present a good representation of this data? Explain.



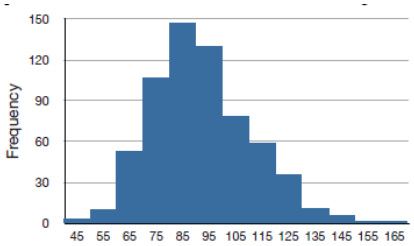
12. Match each description to one of the graphs.
    a. Normal distribution
    b. Positive or right skewed
    c. Negative or left skewed
    d. Bimodal



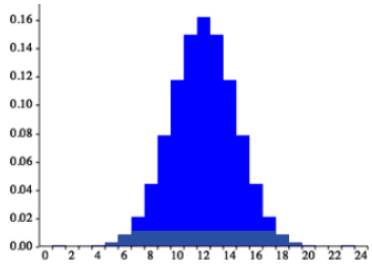
The frequency of times between eruptions of the Old Faithful geyser.



Scores on a 20-point statistics quiz.



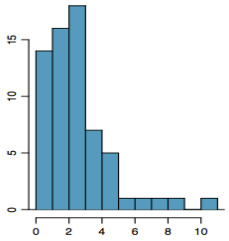
Distribution of scores on a psychology test.



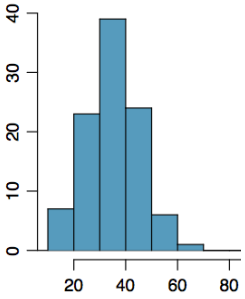
The number of heads in 24 sets of 100 coin flips.

13. Write a sentence or two to describe each distribution in terms of modality, symmetry, skewness and outliers.

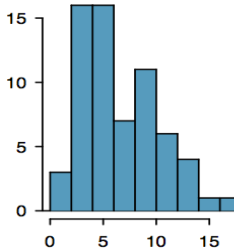
a.



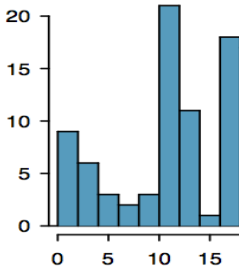
b.



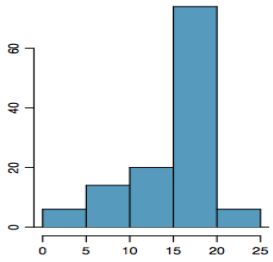
c.



d.



e.



Section 3.3 Summary Statistics: Measures of Center

Calculating Summary Statistics

In addition to graphical and verbal descriptions, we can use numbers to summarize quantitative distributions. We want to know what an "average" value is (where the data is centered), and how spread out the values are. Together, the center and spread provide important information which can be used estimate our population parameters. In this section we will talk about the measures of center and in the next section we will talk about the measures of spread.

Measures of Center

There are a few different types of "averages" that measure the center, and the one we use will depend on the shape of the distribution. We will mention the mode but focus mainly on the two most common "averages": the **mean** and the **median**.

Mode

In the previous section, we saw that the **modes** are related to the peaks where similar values are grouped. A mode is the value where a peak occurs. One way to calculate the mode(s) is to take the midpoint of each peak in the histogram.

Mean

The **mean**, or more formally the arithmetic mean, is what probably comes to mind when you hear the word average. The calculation of the mean uses every data value in the distribution and is therefore strongly affected by skew and outliers.

To calculate the mean of a distribution, we divide the sum of the data values by the number of data values we have. The sample mean is usually represented by $\bar{x}$, a lower-case x with a bar over it, read x-bar. The lower-case letter n is used to represent the number of data values or **sample size**.

Mean

$$\bar{x} = \frac{\text{sum of data values}}{n}$$

Example 1: Mirabel's exam scores for her last math class were: 79, 86, 82, 94. What is her mean test score?

To find the mean test score we need to find the sum of her test scores, then divide the sum by the number of test scores ($n = 4$). The mean is:

$$\bar{x} = \frac{79 + 86 + 82 + 94}{4}$$
$$= 85.25 \text{ points}$$

We will round the sample mean to one more decimal place than the original data. In this case, we would round 85.25 to 85.3 points. Also notice that the mean has the same units as the data and it is important to label it.

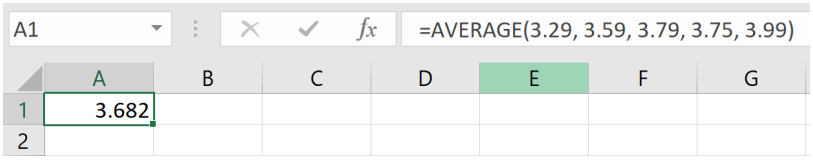
It is reasonable to calculate the mean by hand when the data set is small, but if the data set is large, or if you will be finding additional statistics, then technology is the way to go. We can find the mean of a data set using the spreadsheet formula =AVERAGE.

Example 2: The price of peanut butter at 5 stores was $3.29, $3.59, $3.79, $3.75, and $3.99. Find the mean price using a spreadsheet.

There are two ways to use the =AVERAGE formula. If your data set is not too large, you can enter each value directly into the formula. Using this method, we write

=AVERAGE(3.29 3.59, 3.79, 3.75, 3.99)

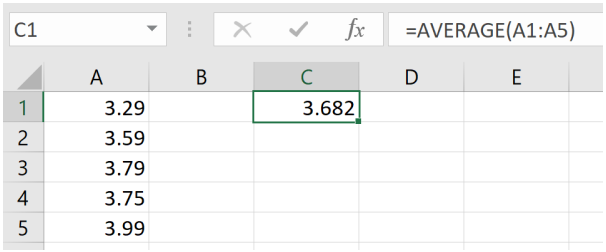
=$3.68



The other method is to enter the data values into a single column (or row) of the spreadsheet and reference the column (or row) range in the formula. We can enter the range by highlighting the data values. As illustrated below, if we enter the data into column A, the formula is

=AVERAGE(A1:A5)

=$3.68



Sometimes when there is a lot of data with repeated values we are given a frequency table.

Example 3: One hundred families from a particular neighborhood are randomly selected and asked to give their annual household income rounded to the nearest $5,000. The results are shown in the frequency table below.

Income (thousands of dollars)	Frequency
$15	6
$20	8
$25	11
$30	17
$35	19
$40	20
$45	12
$50	7

Calculating the mean by hand could get tedious if we try to type in all 100 values:

$$\bar{x} = \frac{\overbrace{15+\cdots+15}^{6 \text{ terms}} + \overbrace{20+\cdots+20}^{8 \text{ terms}} + \overbrace{25+\cdots+25}^{11 \text{ terms}} + \cdots}{100}$$

We could calculate this more easily by noticing that adding 15 to itself six times is the same as $(15)(6) = 90$. Using this simplification, we get

$$\bar{x} = \frac{(15)(6)+(20)(8)+(25)(11)+(30)(17)+(35)(19)+(40)(20)+(45)(12)+(50)(7)}{100}$$

$$= \frac{3390}{100}$$

$$= 33.9$$

The mean household income of our sample is 33.9 thousand dollars, or $33,900.

We could also use =AVERAGE to find the mean for this example, but it would require entering each repeated value individually. If the mean is all we need, then taking advantage of multiplication as repeated addition is the more straightforward way to go. We could also enter the frequency table and the calculation above in a spreadsheet.

Example 4: Extending the last example, suppose a new family moves into the neighborhood and has a household income of $5 million ($5000 thousand). Adding this to our sample, our mean becomes:

$$\bar{x} = \frac{(15)(6)+(20)(8)+(25)(11)+(30)(17)+(35)(19)+(40)(20)+(45)(12)+(50)(7)+(5000)(1)}{101}$$

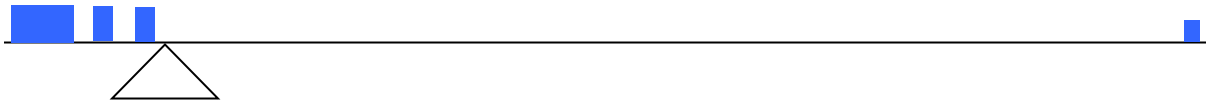
$$= \frac{8390}{101}$$

$$= 83.069$$

While 83.1 thousand dollars, or $83,100 is the correct mean household income for the new sample, it is no longer representative of the neighborhood – in fact, it is greater than every income in the sample aside from the new one we added!

Imagine the data values on a see-saw or balance scale. The mean is the value at the tip of the triangle that keeps the data in balance, like in the picture below.

If we graph our household data, the $5 million value is so far out to the right that the mean has to adjust to keep things in balance.



For this reason, when working with data that is skewed or has outliers, it is common to use a different measure of center, the median.

Median

The **median** of a data set is the "middle" value, when the data are listed in order from smallest to largest. We can also think of the median as the value that has 50% of the data below it and 50% of data above it. As we will discover later, this also makes the median what we call the **50th percentile**.

Median

If the number of data values is odd, then the median is the middle data value
If the number of data values is even, then the median is the mean of the middle pair

Example 5 (odd number of values): Find the median of these quiz scores: 5, 10, 8, 6, 4, 8, 2, 5, 7, 7, 6

We must start by listing the data in order: 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10.

It is helpful to mark or cross off the numbers as you list them to make sure you don't miss any. Also, be sure to count the number of data values in your ordered list to make sure it matches the number of data values in the original list.

In this example there are 11 quiz scores. When the distribution contains an odd number of data values there will be a single number in the middle and that is the median. For small data sets, we can "walk" one value at a time from the ends of the ordered list towards the center to find the median

Lower Half: 2, 4, 5, 5, 6 | Median: 6 | Upper Half: 7, 7, 8, 8, 10

The median test score is 6 points.

Example 6 (even number of values): Now suppose we add another quiz score to the list. Suppose someone in the class got a perfect score of 20 on this very difficult quiz.

Then the ordered list would be: 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10, 20.

There are now 12 quiz scores in our sample. When the distribution contains an even number of data values there will be a pair of values in the middle rather than a single value. Then we take the average of the middle two values.

$$\overbrace{2,4,5,5,6}^{\text{Lower Half}}\ \overbrace{6,7}^{\text{Middle Pair}}\ \overbrace{7,8,8,10,20}^{\text{Upper Half}}$$

$$\text{Median} = \frac{6+7}{2}$$
$$= 6.5 \text{ points}$$

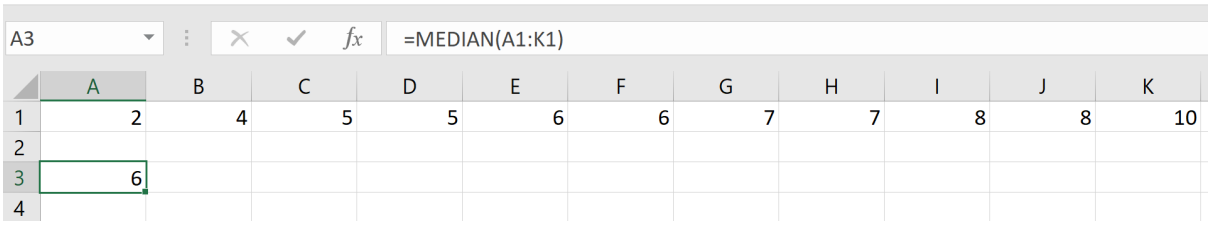
What is important to notice is that despite adding an outlier to our data set, the median is largely unaffected. The median quiz score for the new distribution is 6.5 points.

We can also find the median using the spreadsheet formula =MEDIAN. Just like the spreadsheet function =AVERAGE, we can either list the individual data values in the formula, or we can enter the data values into a row (or column) and use the row range (or column range) in the formula.

Using the data values of the original distribution, we can write function as
=MEDIAN(2, 4, 5, 6, 6, 7, 7, 8, 8, 10) or

=MEDIAN(A1:AK)

=6 points



Example 7: Let's continue with our peanut butter example and find the median both by hand and with a spreadsheet. The price of peanut butter at 5 stores was $3.29, $3.59, $3.79, $3.75, and $3.99.

To find the median by hand, we must first list the prices in order. This give us: $3.29, $3.59, $3.75, $3.79, $3.99

Since there are an odd number of data values in the sample ($n = 5$), we know that the median will be the single data value in the middle of the ordered list.

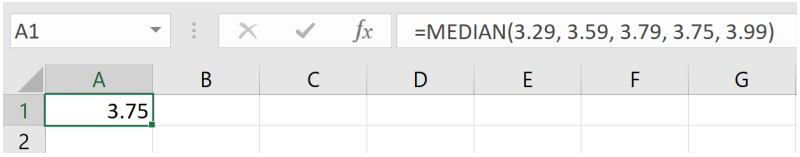
$$\overbrace{3.29,3.59}^{\text{Lower Half}}\ \overbrace{3.75}^{\text{Median}}\ \overbrace{3.79,3.99}^{\text{Upper Half}}$$

The median price of peanut butter at these five stores is $3.75.

Using a spreadsheet, we write

=MEDIAN(3.29, 3.59, 3.79, 3.75, 3.99)

=$3.75



It is worth noting that when you use a spreadsheet to find the median you do not have to order the data first. You can enter the data values in the order they are given to you.

The Relationship Between the Mean and the Median

If a distribution is skewed, the mean is pulled in the direction of the skew, as we saw in the see-saw diagram. In a right skewed distribution, the mean is greater than the median, while in a left skewed distribution, the mean is less than the median. If the distribution is symmetric, the mean and the median will be approximately equal.

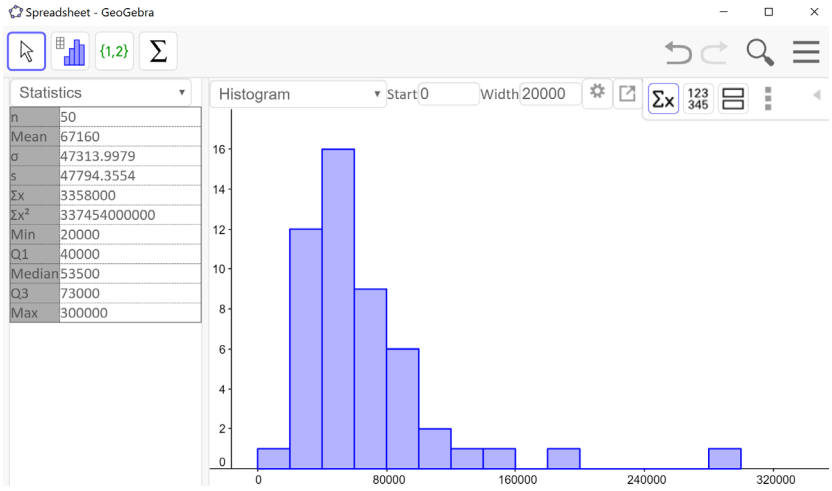
To demonstrate this, we have entered some data in GeoGebra, as previously explained, and made histograms. To see the statistics that GeoGebra calculates, we click on the summation symbol $\left(\sum x\right)$ on the right-hand menu bar.

Example 8: Fifty people from the Portland Metro area who are employed full time were sampled and their annual salaries were recorded (to the nearest thousand dollars). The histogram and summary statistics from GeoGebra are shown below.

From the histogram we can see that the shape of the distribution is unimodal and skewed to the right. We can see from the statistics output on the left that the mean is greater than the median. This is because the few people with higher incomes bring the average up.

$$\text{Mean} = \$67{,}160 \qquad \text{Median} = \$53{,}500$$

$$\text{Mean} > \text{Median}$$

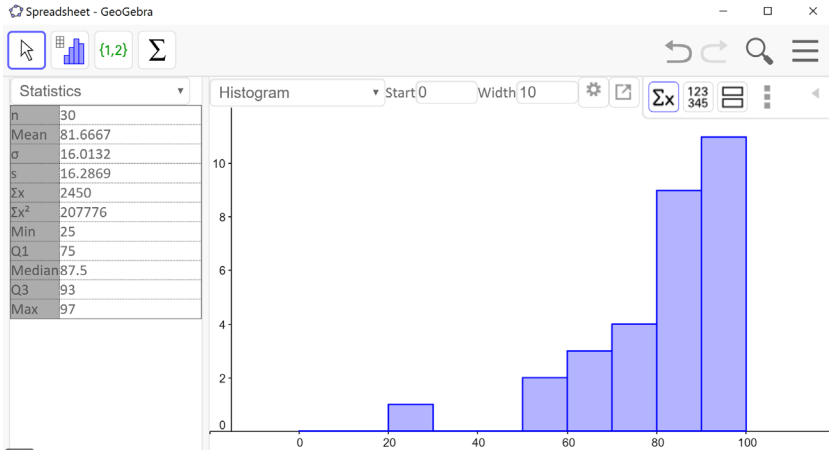


Example 9: A random selection of 30 math 105 exams at PCC were sampled and their scores were recorded. The histogram of the resulting distribution is shown below.

The shape of the distribution is unimodal and skewed to the left. There also appears to be an outlier between 20 and 30. We can see from the statistics output that the mean is less than the median. This is because the low test score brought the average down.

$$\text{Mean} = 81.7 \text{ points} \qquad \text{Median} = 87.5 \text{ points}$$

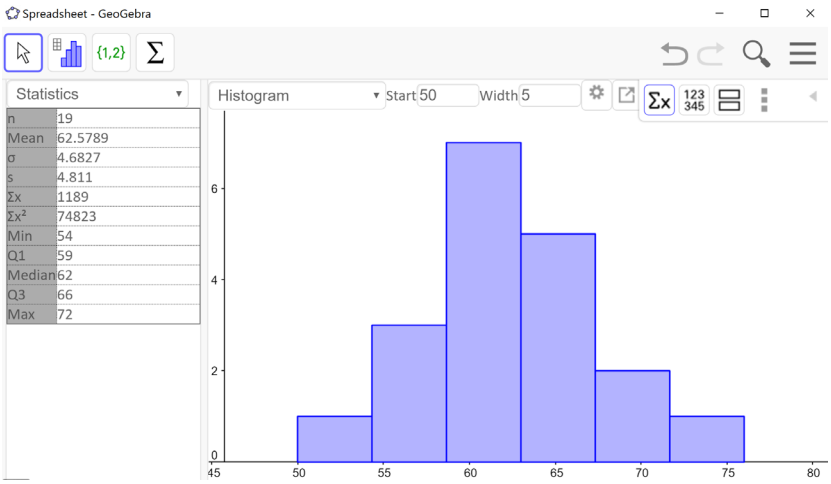
$$\text{Mean} < \text{Median}$$



Example 10: Nineteen people identifying as female were sampled and their heights (in inches) were recorded. The histogram of the resulting distribution is shown below.

The shape of the distribution is unimodal and roughly symmetric. We can also see from the statistics output that the mean and the median are approximately equal.

$$\text{Mean} = 62.6 \text{ inches} \qquad \text{Median} = 62 \text{ inches}$$
$$\text{Mean} \approx \text{Median}$$



We can use these observations in reverse as well. If we know the mean is greater than the median, then we can expect the distribution to be skewed to the right. If the mean is less than the median, then we can expect the distribution to be skewed to the left. When the mean and the median are approximately equal, the distribution is likely to be symmetric.

Example 11: Recent college graduates were asked how much student loan debt they have. The data has a mean of $46,265 and a median of $33,652. Just based on this information, do you expect the distribution to be symmetric, skewed to the left, or skewed to the right?

Since the mean is greater than the median, we can expect the distribution to be skewed to the right.

Exercises 3.3

1. A group of diners were asked how much they would pay for a meal. Their responses were: $7.50, $25.00, $10.00, $10.00, $7.50, $8.25, $9.00, $5.00, $15.00, $8.00, $7.25, $7.50, $8.00, $7.00. $12.00.
    a. Find the mean, including units.
    b. Find the median, including units.
    c. Based on the mean and the median, would you expect the distribution to be symmetric, skewed left, or skewed right? Explain.

2. You recorded the time in seconds it took for 8 participants to solve a puzzle. The times were: 15.2, 18.8, 19.3, 19.7, 20.2, 21.8, 22.1, 29.4.

a. Find the mean, including units.
b. Find the median, including units.
c. Based on the mean and the median, would you expect the distribution to be symmetric, skewed left, or skewed right? Explain.

3. Use the following table is the cost of purchasing a car at a local dealership. Some of the cars sold were new and some were used.
   a. Calculate find the mean, including units.
   b. Can you figure out how to find the median using the frequency table? See if you can do it without listing out all the data values.
   c. Based on the mean and the median, would you expect the distribution to be symmetric skewed left or skewed right? Explain.

Cost (Thousands of dollars)	Frequency
15	3
20	7
25	10
30	15
35	13
40	11
45	9
50	7

4. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer drug is currently under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout their treatment. The following data (in months) are collected.
   a. Find the mean and median of each group.
   b. Compare and contrast the two groups.

   Researcher 1: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

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

5. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the average number of pieces correctly remembered from three chess positions.
   a. Make a histogram for each group.
   b. Find the mean of each group.

c. Find the median of each group.
d. Compare the shapes of the distributions as well as the centers of the three groups.

Non-players	Beginners	Tournament Players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

6. There is evidence that smiling can attenuate judgments of possible wrongdoing. This phenomenon termed the "smile-leniency effect" was the focus of a study by Marianne LaFrance & Marvin Hecht in 1995[6]. The following data are measurements of how lenient the sentences were for three different types of smiles and one neutral control. A higher number indicates greater leniency. The same subject was used for all of the conditions so that may affect the results.
    a. Make a histogram for each smile type and the neutral control.
    b. Find the mean for each type of smile and the neutral control.
    c. Find the median for each type of smile and the neutral control.
    d. Compare the shapes of the distributions as well as the centers for each type of smile and control.

    (The data continues on the next page)

False Smile	Felt Smile	Miserable Smile	Neutral Control
2.5	7	5.5	2
5.5	3	4	4
6.5	6	4	4
3.5	4.5	5	3
3	3.5	6	6
3.5	4	3.5	4.5
6	3	3.5	2
5	3	3.5	6
4	3.5	4	3
4.5	4.5	5.5	3

[6] LaFrance, M., & Hecht, M. A. (1995) Why smiles generate leniency. Personality and Social Psychology Bulletin, 21, 207-214. Adapted from www.onlinestatbook.com, by David M. Lane, et al, used under CC-BY-SA 3.0.

5	7	5.5	4.5
5.5	5	4.5	8
3.5	5	2.5	4
6	7.5	5.5	5
6.5	2.5	4.5	3.5
3	5	3	4.5
8	5.5	3.5	6.5
6.5	5.5	8	3.5
8	5	5	4.5
6	4	7.5	4.5
6	5	8	2.5
3	6.5	4	2.5
7	6.5	5.5	4.5
8	7	6.5	2.5
4	3.5	5	6
3	5	4	6
2.5	3.5	3	2
8	9	5	4
4.5	2.5	4	5.5
5.5	8.5	4	4
7.5	3.5	6	2.5
6	4.5	8	2.5
9	3.5	4.5	3
6.5	4.5	5.5	6.5

7. Make up three data sets with 5 values each that have:
    a. The same mean but different medians.
    b. The same median but different means.

8. The frequency table below shows the number of women's shoes that were sold in an hour at a local shoe store.
    a. Would you treat this data as categorical or quantitative?
    b. How would the bar graph be different from the histogram?
    c. Treat the data as quantitative and find the mean and the median. Are these useful statistics?

Shoe Size	Frequency
5	4
6	4
7	6
8	6
9	5

Section 3.4 Summary Statistics: Measures of Variation

Measures of Variation

Consider these three sets of quiz scores for a 10-point quiz:

Section A: 5 5 5 5 5 5 5 5 5 5

Section B: 0 0 0 0 0 10 10 10 10 10

Section C: 4 4 4 5 5 5 5 6 6 6

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

Thus, in addition to the mean and median, which are measures of center or the "average" value, we also need a measure of how "spread out" or varied each data set is.

There are several ways to measure the variation of a distribution. In this section we will look at the **standard deviation**, **range** and the **interquartile range (IQR)**.

Standard Deviation

The **sample standard deviation, s,** is a measure of variation that tells us how far, on average, the data values deviate, or are different from, the mean. The mean and standard deviation are paired to provide a measure of center and spread for symmetric distributions.

Sample Standard Deviation

$$s = \sqrt{\frac{\text{Sum of the squared deviations from the mean}}{n-1}}$$

where n is the sample size, or the number of data values

We will go through the whole process for calculating the standard deviation. Let's say there is another section of quiz scores:

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

The mean quiz score, like Sections A, B and C, is 5 points.

The first step in finding the standard deviation is to find the deviation, or difference, of each data value from the mean. We will do this in a table. You could also use a spreadsheet to do these calculations.

Data Value	Deviation: (Data Value – Mean)
0	$0 - 5 = -5$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
10	$10 - 5 = 5$

We would like to get an idea of the "average" deviation from the mean, but if we find the average of the values in the second column, the negative and positive values cancel each other out (this will always happen), so to prevent this we square the deviations.

Data Value	Deviation: (Data Value –Mean)	Deviation Squared
0	$0 - 5 = -5$	$(-5)^2 = 25$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
10	$10 - 5 = 5$	$5^2 = 25$

Next, we add the squared deviations and we get

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

Ordinarily, we would then divide by the number of scores, *n*, (in this case, 10) to find the mean of the deviations, but the division by *n* is only done if the data set represents a population. When the data set represents a sample (as it almost always does), we instead divide by $n - 1$ (in this case, $10 - 1 = 9$).

We assume Section D represents a sample, so we will divide by 9. Note that our units are now points-squared since we squared all of the deviations. It is much more meaningful to use the units we started with, so to convert back to points we take the square root.

The sample standard deviation for Section D is

$$s = \sqrt{\frac{50}{9}}$$
$$\approx 2.36 \text{ points}$$

For comparison, here is the standard deviation for each section listed above:

$s_A = 0$ points
$s_B = 5.27$ points
$s_C = 0.82$ points
$s_D = 2.36$ points

For the standard deviation, we usually use two more decimal places than the original data. This tells us that on average, scores were 2.36 points away from the mean of 5 points. In summary, here are the steps to calculate the standard deviation by hand.

Calculating the Sample Standard Deviation

1. Find the deviations by subtracting the mean from each data value
2. Square each deviation
3. Add the squared deviations
4. Compute the square root of the sum divided by $n-1$:

$$s = \sqrt{\frac{\text{Sum of the squared deviations from the mean}}{n-1}}$$

There are a few important characteristics we want to keep in mind when finding and interpreting the standard deviation.

- The standard deviation is never negative. It will be zero if all the data values are equal and get larger as the data spreads out.
- The standard deviation has the same units as the original data and it is important to label it.
- The standard deviation, like the mean, can be highly influenced by outliers.

Example 1: To continue our peanut butter example, we will find the standard deviation of this sample: \$3.29, \$3.59, \$3.79, \$3.75, and \$3.99.

The first thing we need to find is the sample mean, and we know it is \$3.68 from our previous work. Next, we need to find the deviation from the mean for each data value and square it.

Data Value	Deviation	Deviation Squared
\$3.29	3.29 – 3.68 = -0.39	$(-0.39)^2 = 0.1521$
\$3.59	3.59 – 3.68 = -0.09	$(-0.09)^2 = 0.0081$
\$3.79	3.79 – 3.68 = 0.11	$(0.11)^2 = 0.0121$
\$3.75	3.75 – 3.68 = 0.07	$(0.07)^2 = 0.0049$
\$3.99	3.99 – 0.68 = 0.31	$(0.31)^2 = 0.0961$

The sum of the deviations squared is

$0.1521 + 0.0081 + 0.0121 + 0.0049 + 0.0961 = 0.2733$ dollars-squared.

The sample standard deviation is

$$s = \sqrt{\frac{0.2733}{4}}$$
$$\approx \$0.2614$$

Since the units are dollars, we will round to two decimal places rather than two more than the data. This gives us a standard deviation of $0.26. Together with the mean this tells us that on average, the cost of a jar of peanut butter is $0.26 away from the mean of $3.68.

Calculating the standard deviation by hand can be quite a nuisance when we are dealing with a large data set, so we can also use technology. We use the spreadsheet function =STDEV.S to find the *sample* standard deviation. Notice that this is different from the population standard deviation, which uses the function =STDEV.P.

Just like the spreadsheet functions =AVERAGE and =MEDIAN, we can either list the individual data values in the formula, or we can enter the data values into a row or column and use the row or column range in the formula.

Example 2: The total cost of textbooks for the term was collected from 36 students. Use a spreadsheet to find the mean, median, and standard deviation of the sample.

$140	$160	$160	$165	$180	$220	$235	$240	$250
$260	$280	$285	$285	$285	$290	$300	$300	$305
$310	$310	$315	$315	$320	$320	$330	$340	$345
$350	$355	$360	$360	$380	$395	$420	$460	$460

Since we are finding more than one statistic for this data set, it is much more efficient to enter the data values into a row or column and reference the range in each of the formulas. Entering the data in column A, the formulas are:

Mean: =AVERAGE(A1:A36)

= $299.58

Median: =MEDIAN(A1:A36)

= $307.50

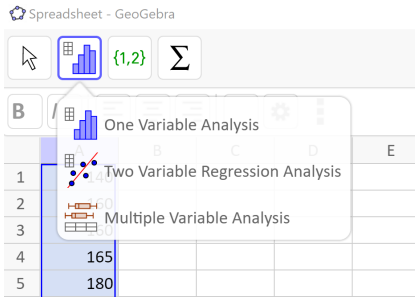
Standard Deviation: =STDEV.S(A1:A36)

= $78.68

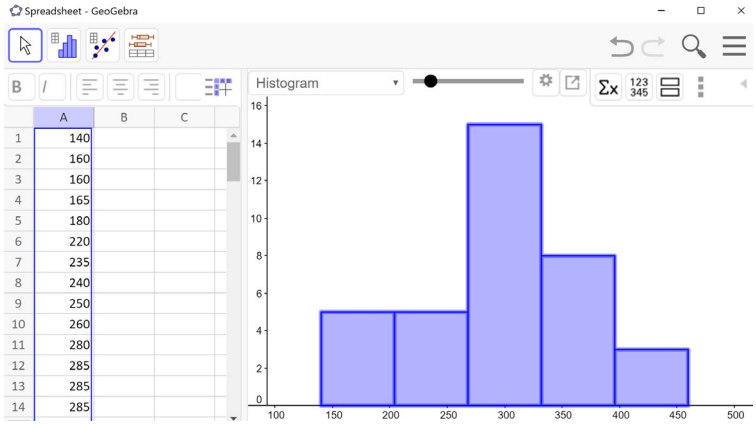
The mean and the median are relatively close to each other, so we can expect the distribution to be approximately symmetric with maybe a slight skew to the left, since the mean is smaller. The mean and standard deviation together tell us that the average cost of textbooks for a term is about $299.58, give or take $78.68.

In addition to a spreadsheet, we will continue our use of GeoGebra. Let's take a look at how to use GeoGeogebra to find the mean, median and standard deviation for the last example. We begin just like we did for making a histogram.

Example 2 Continued: We enter the textbook data into column A of the spreadsheet in GeoGebra. (Main Menu →View →Spreadsheet). Next, select the column title of your data, click on the histogram in the menu bar on the left, and select One Variable Analysis.

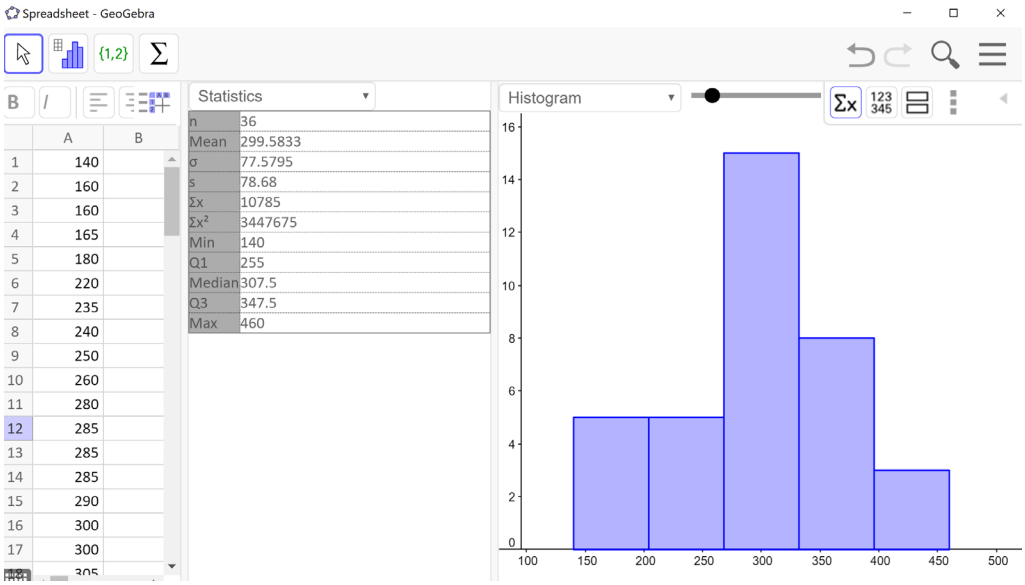


Then you will see the histogram. It is always a good idea to check the shape of your distribution before calculating anything. Notice our histogram matches what we found about the distribution from the mean and median. It is approximately symmetric or slightly skewed to the left.



Next, we click the summation symbol $\left(\sum x\right)$ in the menu bar on the right. The list of summary statistics will pop up as you can see in the image below. We see that the mean is \$299.58, the median is \$307.50 and the sample standard deviation is \$78.68 – just like we found using the spreadsheet formulas.

The statistics we will use are the sample size (n), the mean, median, and the sample standard deviation (s). The last five entries in the table – min, Q1, median, Q3, and max – together make up the 5-number summary which we will learn about shortly!



The standard deviation is the measure of variation that we pair with the mean for approximately symmetric distributions. This pairing should make sense because the standard deviation uses the mean in its calculation. But what about the median? What measure of variation do we pair with it?

Range

One candidate is the **range**. The range tells us the spread or width of the entire data set. We calculate the range as the difference between the maximum and minimum value.

Range

Range = Max – Min

However, the range is not a very good measure of variation since it is very strongly affected by skew and outliers. Consider, for example, the distribution of full time salaries in the United States. Many people earn a minimum wage salary, while others like Jeff Bezos (Amazon) and Bill Gates (Microsoft) earn millions (if not billions!). A range this large does very little to help us get a sense of the spread where most of the data values lie.

Quartiles and the Interquartile Range

Instead, the measure of variation that we pair with the median is the **interquartile range (IQR)**. The IQR tells us the width of the middle 50% of data values. By cutting off the lower and upper 25% of data values, we are able to ignore extreme values and provide a more accurate sense of how spread out the distribution is.

The IQR is calculated as the difference between the third quartile (Q_3) and the first quartile (Q_1). Before we can calculate the interquartile range, though, we need to learn how to find the first and third quartiles.

Interquartile Range (IQR)

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

As the name implies, **quartiles** are values that divide the data into quarters. The **first quartile (Q_1)** is the value that 25% of the data lie below. The **third quartile (Q_3)** is the value that 75% of the data lie below. As you might have guessed, the second quartile is the same as the median since 50% of the data values lie below it.

We have seen that the data is split in half by the median so to split it into quarters, we find the median of each half of the data.

Quartiles

Q_1 the median of the lower half of the data
Q_2 is the median of the whole data set
Q_3 is the median of the upper half of the data

If there is an odd number of data values, we don't use the median in either half

Example 3 (even number of data values): Suppose we have measured the height, in inches, of 12 people who identify as female. The data values are listed below. Find the interquartile range.

59 60 69 64 70 72 66 64 67 66 63 61

Just like when finding the median, we must first order the data.

59 60 61 63 64 64 66 66 67 69 70 72

Then we divide the data into two halves. In the case of an even sample size, we split the distribution down the middle. The first 6 data values are the lower half and the next 6 data values are the upper half. Then we find the median of each. The median of the lower half is Q_1 and the median of the upper half is Q_3.

$$\overbrace{59, 60, \underbrace{61, 63}_{Q_1=\frac{61+63}{2}=62}, 64, 64}^{\text{Lower Half}} \overbrace{66, 66, \underbrace{67, 69}_{Q_3=\frac{67+69}{2}=68}, 70, 72}^{\text{Upper Half}}$$

In this data, $Q_1 = 62$ inches and $Q_3 = 68$ inches. Then we subtract to find the IQR.

$$\begin{aligned}\text{IQR} &= Q_3 - Q_1 \\ &= 68 - 62 \\ &= 6 \text{ inches}\end{aligned}$$

This tells us the that the middle 50% of the women's heights lie within an interval of 6 inches.

Example 4 (odd number of data values): Suppose we added one more height (68 inches) to the data set from the previous example. We will again find the interquartile range of the heights.

59 60 61 63 64 64 66 66 67 68 69 70 72

The data are already in order, but we have an odd number of values. To deal with this we do not use the median in the upper or lower halves. The lower half will include the values strictly below the median, and the upper half will include the values strictly above the median.

Lower Half: $59, 60, 61, 63, 64, 64$ — $Q_1 = \frac{61+63}{2} = 62$

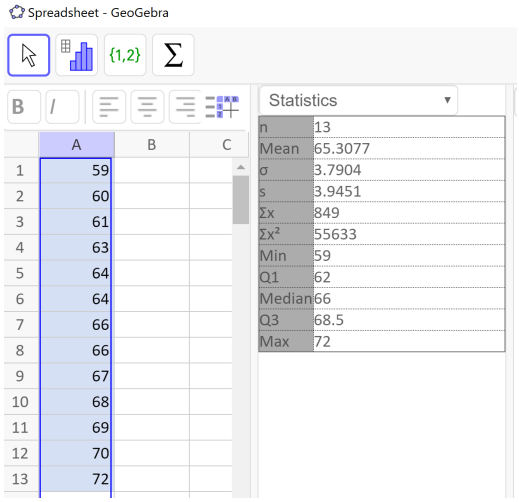
Median (Ignore): 66

Upper Half: $66, 67, 68, 69, 70, 72$ — $Q_3 = \frac{68+69}{2} = 68.5$

Then the interquartile range is:

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 68.5 - 62 \\ &= 6.5 \text{ inches} \end{aligned}$$

If the data set is small, we can find the first and third quartiles by hand, but we have also seen that they are part of the output from GeoGebra. Here is the output for this data.



From the list of summary statistics we can see that $Q_1 = 62$ and $Q_3 = 68.5$ inches. Now we can calculate the interquartile range.

$$\begin{aligned} \text{IQR} &= 68.5 - 62 \\ &= 6.5 \text{ inches} \end{aligned}$$

This is the same value we found by hand.

It is important to note that we are not using spreadsheets for the five-number summary because they do not calculate the quartiles in the same way, so they will not give the same results. Now that we have learned how to find the quartiles we can make a five-number summary and boxplot.

The Five-Number Summary and Boxplots

The **five-number summary** is made up of the minimum, Q_1, median, Q_3, and the maximum. These five values divide the data into quarters. A **boxplot**, also called a **box-and-whisker plot**, is a graphical representation of the five-number summary. Each region of the boxplot contains approximately the same number of data values, so we can see the spread for each region. We can find the five-number summary and draw a boxplot by hand or by using GeoGebra. In our last example the five-number summary from GeoGebra is: 59, 62, 66, 68.5, 72 inches.

Five-Number Summary

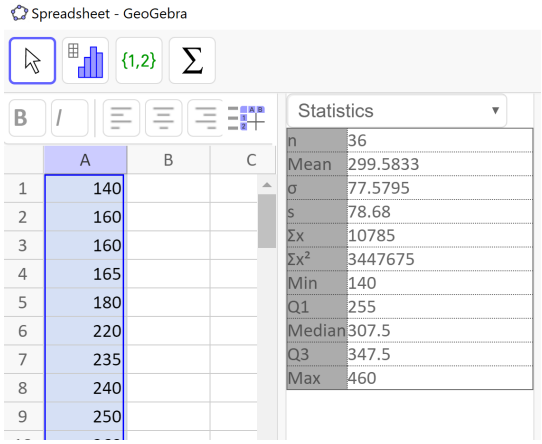
Minimum, Q_1, Median, Q_3, Maximum

We will use GeoGebra to find the five-number summary for the next example and then explain how to draw a boxplot.

Example 5: Let's continue with the cost of textbook data from Example 2. Use GeoGebra to find the five-number summary for this sample and draw a boxplot by hand.

$140	$160	$160	$165	$180	$220	$235	$240	$250
$260	$280	$285	$285	$285	$290	$300	$300	$305
$310	$310	$315	$315	$320	$320	$330	$340	$345
$350	$355	$360	$360	$380	$395	$420	$460	$460

As we found before, here is the GeoGebra output. The last five entries of the summary statistics are the five-number summary. Remember to label all of your statistics with units.

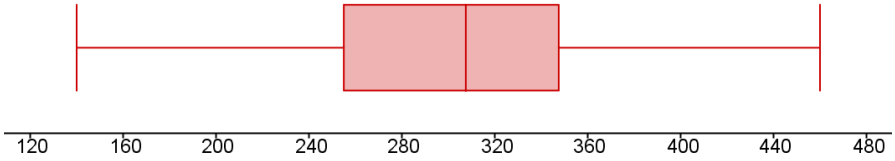


The five-number summary is

Min	Q_1	Median	Q_3	Max
\$140	\$255	\$307.50	\$347.50	\$460

To draw the boxplot, we will first draw a number line that extends a little beyond the minimum and maximum values, and choose a scale. We decided to draw our number line from \$120 to \$480, in increments of \$40. Then we add a meaningful title and units.

Next, make vertical lines at the first quartile, median and third quartile and connect them to form a box. This is the middle 50% of the data and you might notice that the width of the box is the IQR. Then, extend the "whiskers" out to the minimum and maximum values. Note that a boxplot does not have a vertical scale and the height of the box does not matter. Our boxplot looks like this:



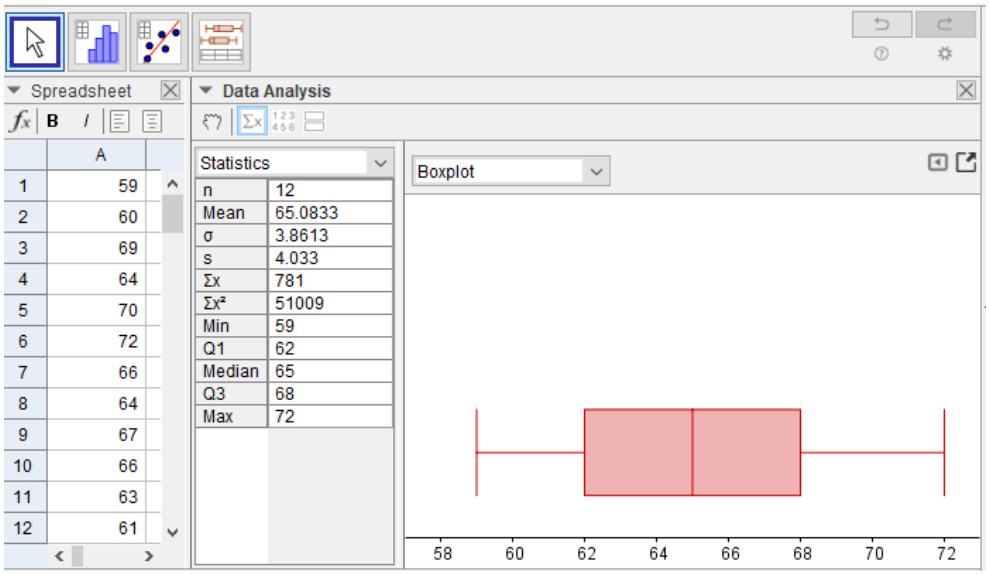
Cost of Textbooks for One Term in Dollars, for a Sample of 36 Students

GeoGebra will also draw boxplots for us. We enter and select the data values like we have done before and select One Variable Statistics. This brings up the graphics window with a histogram by default. Use the drop-down menu to select the boxplot. We also click on $\sum x$ to show the summary statistics.

Example 6: We will continue with our height data from the 12 people who identify as women. Find the five-number summary and create a boxplot using GeoGebra.

59 60 69 64 70 72 66 64 67 66 63 61

Following the steps above we have the following GeoGebra output. The last five entries in the statistics table are the five-number summary and we have the boxplot on the right.

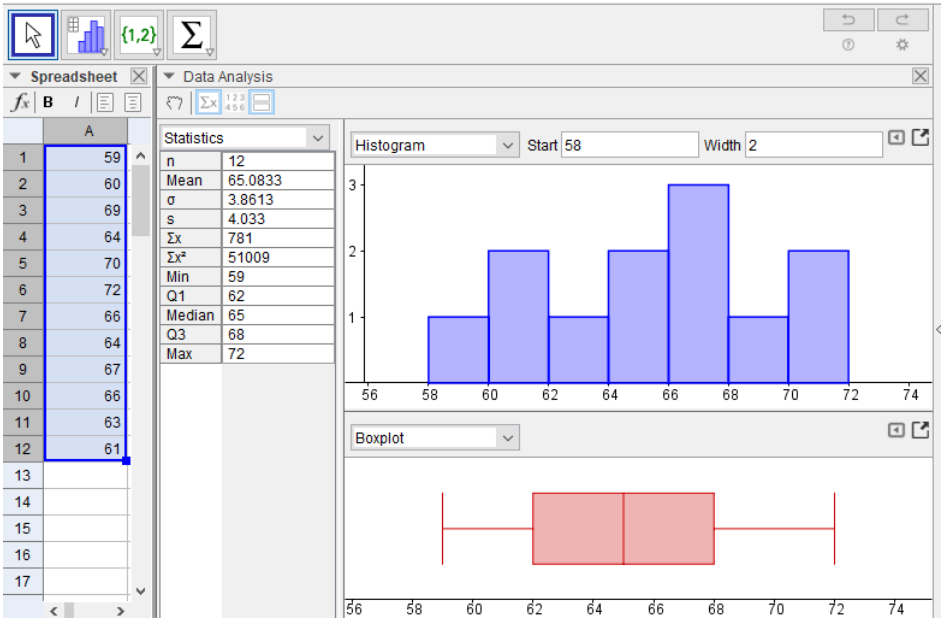


Here is the five-number summary:

Min	Q_1	Median	Q_3	Max
59 in	62 in	65 in	68 in	72 in

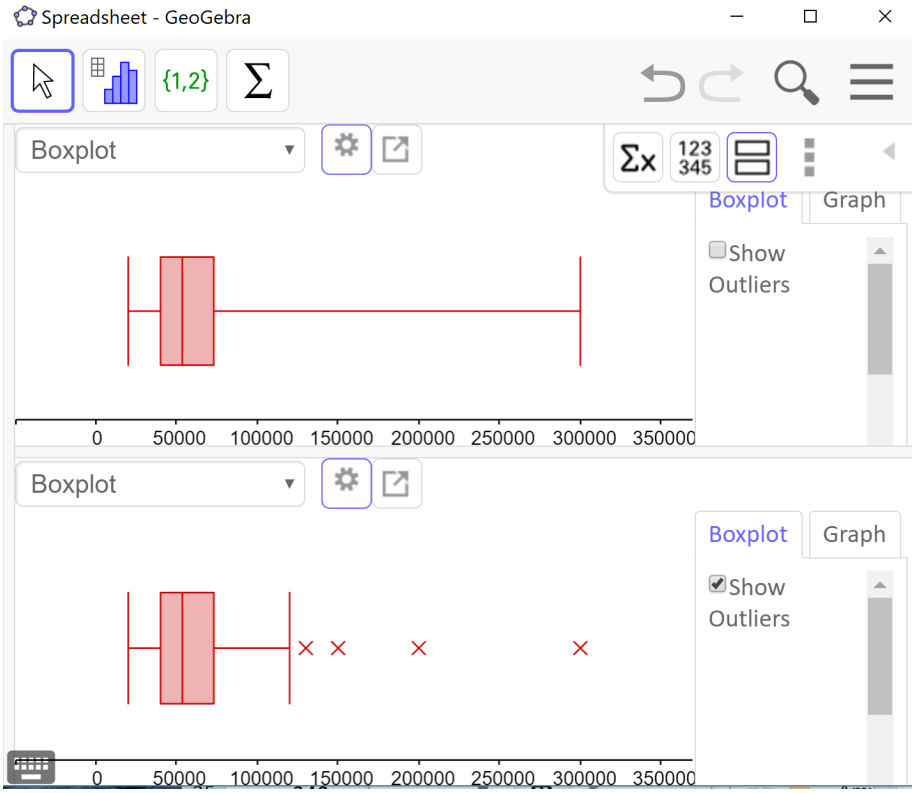
For this data, the two sides of the box and the two whiskers are approximately the same width. This suggests that the distribution is symmetric. We can verify this by noticing that the mean is approximately equal to the median.

The boxplot can tell us the shape of the distribution, but we cannot tell how many peaks the data has. For that we need a histogram. We can see the histogram and boxplot together by selecting the icon that looks like two rectangles stacked, or an = sign.



Now we have a full picture of this data.

The default boxplot in GeoGebra is called a **modified box plot**, which shows the data values that are outliers with an X but requires a few more steps to make by hand. To change from the modified box plot to a regular box plot, click on the left pointing arrow in the boxplot window (downloaded version) or the settings wheel (online version) for options, and uncheck "show outliers." The output window below shows two side-by-side boxplots (the regular boxplot on top and the modified boxplot on the bottom) illustrating the distribution of the annual salaries for 50 randomly selected full-time workers in the Portland Metro area.



From the upper boxplot we can see that this distribution is skewed to the right and the upper quarter of the data is very spread out. It is natural to think of the data values as being evenly spread out in each region, but that is quite often not the case. From the lower boxplot we can see that there are 4 data values that are considered outliers and how far away the last data value is from the others. This is why it is useful to show outliers on a boxplot.

Modified Boxplot (Optional)

If you are curious to know how a modified boxplot is made, we will explain it briefly. There is a rule called the **1.5*IQR** rule to determine which points are considered outliers. An outlier is a point that is more than 1.5 IQRs away from the middle 50% of the data (the box in the boxplot). We know how to calculate the IQR and then we multiply that by 1.5. We subtract that from Q_1 to find the **lower fence** and add that value to Q_3 to find the **upper fence**. Values beyond the fences are considered outliers and are

drawn with an X or a star. Then we draw the whiskers of a modified box plot to the furthest data value inside the fence on each side.

Percentiles

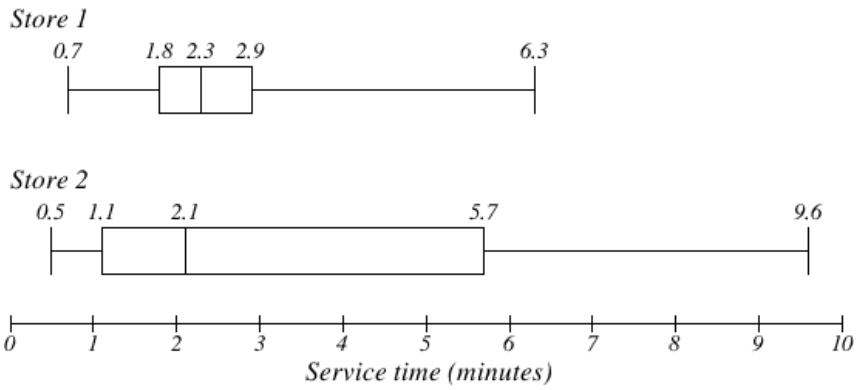
Back when we were finding the median, we mentioned that the median is also called the 50th percentile, because 50% of the data values lie below it. We can define any **percentile** as the data value with that percentage of values below it. Since we have found the quartiles, we can also identify the 25th and 75th percentiles for our data. Q_1 is the 25th percentile because 25% of the data values lie below it and Q_3 is the 75th percentile because 75% of the data values lie below it.

Percentiles are used when comparing the growth of children to the population and in the results of standardized tests, such as the SAT test. If a person scored in the 83rd percentile, that means they scored higher than 83% of the people who took the test.

Comparing Distributions

Box plots and percentiles are particularly useful for comparing data from two populations.

Example 7: The box plots of service times for two fast-food restaurants are shown below. Compare the length of time to get served at the two restaurants. Which one should you go to if you are in a hurry?



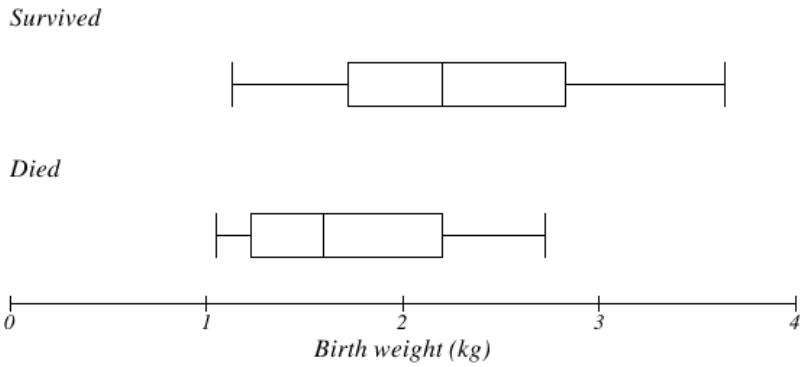
Store 2 has a slightly shorter median service time (2.1 minutes vs. 2.3 minutes), but the service times are less consistent, with a wider spread of the data.

The 75th percentiles are 2.9 and 5.7 minutes. That means 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 opinion about luck – 25% of customers at store 2 had to wait between 5.7 and 9.6 minutes.

Example 8: The boxplots below show the 5-number summaries of the birth weights, in kilograms, of infants with severe idiopathic respiratory distress

syndrome (SIRDS)[7]. The boxplots are separated to show the birth weights of infants who survived and those that did not. What can we conclude from this data?



Comparing the two groups, the boxplot reveals that the birth weights of the infants that died appear to be, overall, smaller than the weights of infants that survived. In fact, we can see that the median birth weight of infants that survived is about the same as the third quartile of the infants that died.

Similarly, we can see that the 25th percentile of the survivors is larger than the 50th percentile of those that died, meaning that over 75% of the survivors had a birth weight larger than the median birth weight of those that died.

Looking at the maximum value for those that died and the third quartile of the survivors, we can see that over 25% of the survivors had birth weights higher than the heaviest infant that died.

The box plots give us a quick, though informal, way to determine that birth weight is quite likely linked to the survival of infants with SIRDS.

Z-Scores

Have you ever heard the saying that you can't compare apples and oranges? It turns out that you can - provided we standardize their measures first!

We will be using the standard score called a Z-score, which is a method commonly used with unimodal and symmetric distributions (called **normal** or **nearly normal distributions**). Z-scores may be used with any data, but if the distribution is skewed, then the distribution of Z-scores will also be skewed.

To calculate the Z-score for a data value, we find out how far away from the mean it is by subtracting. Then we divide by the standard deviation to see how many standard deviations that is. Thus, the **Z-score** of a data value is the number of standard deviations it is away from the mean.

[7] van Vliet, P.K. and Gupta, J.M. (1973) Sodium bicarbonate in idiopathic respiratory distress syndrome. *Arch. Disease in Childhood*, 48, 249–255. As quoted on http://openlearn.open.ac.uk/mod/oucontent/view.php?id=398296&section=1.1.3

Z-score

$$Z = \frac{\text{data value - mean}}{\text{standard deviation}}$$

Be sure to calculate the difference first, then divide

If a data value is above the mean, its Z-score will be positive. If a data value is below the mean, its Z-score will be negative. Therefore, if a data value is one standard deviation above the mean, its Z-score is +1. If it is 2.5 standard deviations below the mean, its Z-score is -2.5. Note that the units of Z-scores are standard deviations, not the units of the data values.

We can use Z-scores to determine the relative unusualness of a data value with respect to its own distribution. That is what allows us to compare two unlike items. The convention in statistics is to say that a data value is **unusual** if it is more than 2 standard deviations from the mean, or in other words, if its Z-score is less than -2 or greater than +2.

Example 9: The oranges at a local grocery store have a mean diameter of 5.8 inches and a standard deviation of 1.2 inches. The apples, on the other hand, have a mean diameter of 4.2 inches and a standard deviation of 0.6 inches.

Ali closes their eyes and selects and apple and an orange. When they look at both pieces of fruit, they seem small. If the orange has a diameter of 4.2 inches and the apple has a diameter of 3.5 inches, which is smaller relative to their respective piles of fruit?

To determine which fruit is relatively smaller, Ali can find each of their Z-scores.

$$Z_{\text{Orange}} = \frac{4.2 - 5.8}{1.2} \qquad Z_{\text{Apple}} = \frac{3.5 - 4.2}{0.6}$$
$$= -1.33 \text{ standard deviations} \qquad = -1.17 \text{ standard deviations}$$

By convention, Z-scores are rounded to two decimal places, so we see that the orange is 1.33 standard deviations below its mean and the apple is 1.17 standard deviation below its mean. The orange is therefore smaller relative to its distribution since its Z-score is less than the apple's Z-score.

We can also see from the Z-scores that neither fruit has an unusually small diameter since each piece of fruit is less than 2 standard deviations from its mean.

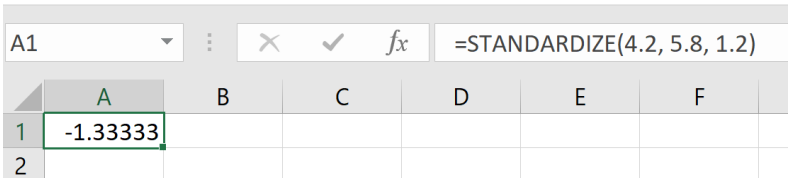
We can also find Z-scores using a spreadsheet with this formula:

=STANDARDIZE(data value, mean, standard deviation)

To verify our apple and orange Z-scores, we would write:

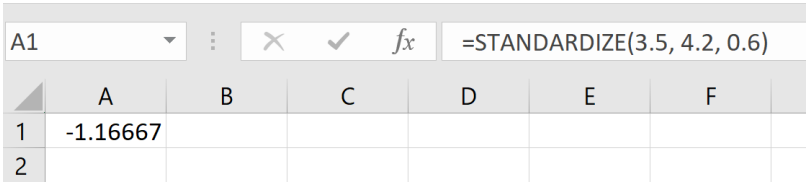
Apple: =STANDARDIZE(4.2, 5.8, 1.2)

= -1.33 standard deviations



Orange: =STANDARDIZE (3.5, 4.2, 0.6)

= -1.17 standard deviations



Example 10: The mean weight of men over the age of 20 is 195.7 pounds[8] with a standard deviation of 29.8 pounds. The mean weight of domestic cats is 8.6 pounds with a standard deviation of 1.2 pounds. (The standard deviation for men's weights is estimated. The cat's mean weight is based on ideal cat weight and the standard deviation is approximate).

At his peak, Andre the Giant, the 7-foot-4-inch French professional wrestler and actor, weighed 520 pounds. When Georgie the cat was at his peak he weighed 24 pounds. Who was more giant – Andre the Giant or Georgie the cat?

Since the weights of cats and men cannot be compared directly, we will need to calculate the Z-scores.

$$Z_{\text{Andre}} = \frac{520 - 195.7}{29.8} = 10.88 \text{ standard deviations}$$

$$Z_{\text{Georgie}} = \frac{24 - 8.6}{1.2} = 12.83 \text{ standard deviations}$$

Using the standardize function, we would write:

Andre: =STANDARDIZE (520, 197.5, 29.8)

= 10.88 standard deviations

Georgie: =STANDARDIZE (24, 8.6, 1.2)

= 12.83 standard deviations

[8] 2016 CDC Report. https://www.cdc.gov/nchs/data/series/sr_03/sr03_039.pdf. The study included all ethnicities but the report does not say whether transgendered men were included.

Since both Z-scores are greater than 2 standard deviations, both weights are extremely unusual. However, since the Z-score for Georgie's weight is larger, he is even more giant than Andre the Giant.

Exercises 3.4

Many of the datasets from Exercises 3.3 are repeated here so you can use your previous work to help you.

1. A group of diners were asked how much they would pay for a meal. Their responses were: $7.50, $25.00, $10.00, $10.00, $7.50, $8.25, $9.00, $5.00, $15.00, $8.00, $7.25, $7.50, $8.00, $7.00. $12.00.
    a. Using your mean from section 3.3, find the standard deviation of this data. Explain what the mean and standard deviation tell you about how much the group of diners would pay for a meal.
    b. Calculate the five-number summary for this data.
    c. Calculate the range and IQR for this data.
    d. Create a boxplot for the data.

2. You recorded the time in seconds it took for 8 participants to solve a puzzle. The times were: 15.2, 18.8, 19.3, 19.7, 20.2, 21.8, 22.1, 29.4.
    a. Using your mean from section 3.3, find the standard deviation of this data. Explain what the mean and standard deviation tell you about how much the group of diners would pay for a meal.
    b. Calculate the five-number summary for this data.
    c. Calculate the range and IQR for this data.
    d. Create a boxplot for the data.

3. Use the following table is the cost of purchasing a car at a local dealership. Some of the cars sold were new and some were used.
    a. Find the standard deviation of this data. Explain what the mean and standard deviation tell you about how much the cars are selling for.
    b. Calculate the five-number summary for this data.
    c. Calculate the range and IQR.
    d. Create a boxplot for the data.

Cost (Thousands of dollars)	Frequency
15	3
20	7
25	10
30	15
35	13
40	11
45	9
50	7

4. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer drug is currently under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout their treatment. The following data (in months) are collected.
    a. Find the standard deviation of each group.
    b. Calculate the 5-number summary for each group.
    c. Calculate the range and IQR for each group.
    d. Create side-by-side boxplots and compare and contrast the two groups.

    Researcher 1: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

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

5. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the average number of pieces correctly remembered from three chess positions.
    a. Find the standard deviation of each group.
    b. Calculate the 5-number summary for each group.
    c. Calculate the range and IQR for each group.
    d. Create side-by-side boxplots and compare and contrast the two groups.

Non-players	Beginners	Tournament Players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

6. There is evidence that smiling can attenuate judgments of possible wrongdoing. This phenomenon termed the "smile-leniency effect" was the focus of a study by Marianne LaFrance & Marvin Hecht in 1995[9]. The following data are measurements of how lenient the sentences were for three different types of smiles and one neutral control.

[9] LaFrance, M., & Hecht, M. A. (1995) Why smiles generate leniency. Personality and Social Psychology Bulletin, 21, 207-214. Adapted from www.onlinestatbook.com, by David M. Lane, et al, used under CC-BY-SA 3.0.

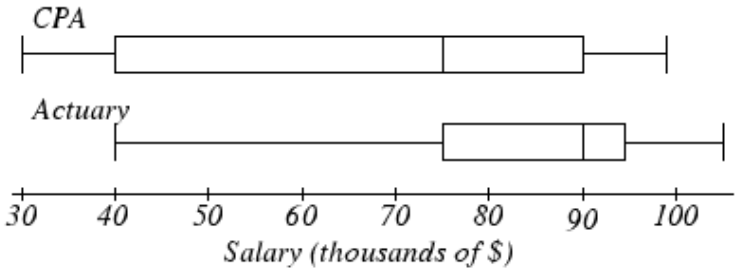
The same subject was used for all of the conditions so that may affect the results. The second column is a continuation of the first column.

a. Find the standard deviation for each type of smile and the neutral control.
b. Calculate the 5-number summary for type of smile and the neutral control.
c. Calculate the range and IQR for each type of smile and the neutral control.
d. Create side-by-side boxplots and compare and contrast the four groups.

False Smile	Felt Smile	Miserable Smile	Neutral Control
2.5	7	5.5	2
5.5	3	4	4
6.5	6	4	4
3.5	4.5	5	3
3	3.5	6	6
3.5	4	3.5	4.5
6	3	3.5	2
5	3	3.5	6
4	3.5	4	3
4.5	4.5	5.5	3
5	7	5.5	4.5
5.5	5	4.5	8
3.5	5	2.5	4
6	7.5	5.5	5
6.5	2.5	4.5	3.5
3	5	3	4.5
8	5.5	3.5	6.5
6.5	5.5	8	3.5
8	5	5	4.5
6	4	7.5	4.5
6	5	8	2.5
3	6.5	4	2.5
7	6.5	5.5	4.5
8	7	6.5	2.5
4	3.5	5	6
3	5	4	6
2.5	3.5	3	2
8	9	5	4
4.5	2.5	4	5.5
5.5	8.5	4	4
7.5	3.5	6	2.5
6	4.5	8	2.5
9	3.5	4.5	3
6.5	4.5	5.5	6.5

7. Make up two data sets with 5 numbers each that have:
    a. The same mean but different standard deviations.
    b. The same standard deviation but different means.

8. The side-by-side boxplots show salaries for actuaries and CPAs.
    a. Estimate the 25th, 50th and 75th percentiles for CPA and actuary salaries.
    b. Deshawn makes the median salary for an actuary. Kelsey makes the first quartile salary for a CPA. Who makes more money? How much more?
    c. What percentage of actuaries make more than the median salary of a CPA?
    d. What percentage of CPAs earn less than all actuaries?



9. Suppose you buy a new car whose advertised gas mileage is 25 mpg (miles per gallon). After driving the car for several months, you find that you are getting only 21.4 mpg. You phone the manufacturer and learn that the standard deviation of gas mileage for cars of that model is 1.15 mpg.
    a. Find the Z-score for the gas mileage of your car.
    b. Does it appear that your car is getting unusually low gas mileage? Explain your answer using your Z-score.

10. This data is a sample of the average number of hours per year that a driver is delayed by road congestion in 11 cities: 56, 53, 53, 50, 46, 45, 44, 43, 42,40, 36
    a. Find the mean and the standard deviation, including units.
    b. What is the Z-score for the city with an average delay time of 42 hours per year?

11. You scored an 89 on a math test where the class mean and standard deviation are 75 points and 7 points respectively. You scored a 65 on an English test where the mean and standard deviation are 53 points and 4 points, respectively. In which class did you do better? Explain your answer using Z-scores.

12. Poe, the Clydesdale horse has a world record breaking height of 20.2 hands. All Clydesdale horses have a mean height of 16.5 hands and a standard deviation of 1.85 hands. The last Great Dane to hold the world record for dog height was Gibson who was 107 cm tall. Great Danes have a mean height of 81 cm and a standard deviation of 13 cm. Which animal is taller compared to their respective breed? Explain your answer using Z-scores.

Chapter 4: Probability

Student Outcomes for this Chapter

Section 4.1: Contingency Tables

Students will be able to:

- Relate Venn diagrams and contingency tables
- Calculate percentages from a contingency table
- Calculate *and* empirical probabilities
- Calculate *or* empirical probabilities
- Calculate conditional probabilities
- Determine whether two characteristics are independent

Section 4.2: Theoretical Probability

Students will be able to:

- Write the sample space for theoretical probability situations
- Identify certain and impossible events
- Calculate the theoretical probability of a complement
- Determine the difference between empirical and theoretical probability
- Explain the Law of Large Numbers
- Identify independent and dependent events
- Calculate *and* theoretical probabilities
- Identify overlapping and disjoint sets
- Calculate *or* theoretical probabilities
- Calculate probability values for simple games

Section 4.3: Expected Value

Students will be able to:

- Make a probability model
- Calculate the expected value for a probability model
- Determine whether a game is fair

Section 4.1 Contingency Tables

When we looked at categorical data in the previous chapter, it was related to a single variable, or characteristic of interest, such as favorite movie or car color. To illustrate the data, we made a frequency table and used it to create a pie chart or bar chart. But what if we want to illustrate the relationship between two categorical variables? To do this, we can use a contingency table.

Contingency Tables

A **contingency table** summarizes all the possible combinations for two categorical variables. Each value in the table represents the number of times a particular combination of outcomes occurs. For example, suppose we randomly select 250 households from the greater Portland area and ask whether they have a cat and whether they have a dog. In this case, "have a cat" and "have a dog" are the two variables, and each variable has two categories: Yes and No.

To create the contingency table, we make columns for the categories of one variable, and rows for the categories of the other variable. We also add a row and column for the subtotals of each category. Each cell of the resulting table contains the number of outcomes having the characteristics of the intersecting row and column categories. For our dog and cat example, the table would look like this:

	Dog	**No Dog**	**Total**
Cat	*Yes Cat and Yes Dog*	*Yes Cat and No Dog*	*Yes Cat Total*
No Cat	*No Cat and Yes Dog*	*No Cat and No Dog*	*No Cat Total*
Total	*Yes Dog Total*	*No Dog Total*	*Grand Total*

Suppose that of the 250 households surveyed, 180 said they have a cat, 95 said they have a dog, and 52 said they have both a cat and a dog. We can use this information to fill in the cells of the table.

The first cell we can fill in is the **grand total**, which is the total number of subjects in the study. In this case, there are 250 households participating in the survey. The next two cells we can fill in are the total number of households that have a cat, 180, and the total number of households that have a dog, 95. The final cell we can fill in from the given information is the intersection of the having a dog column and a having a cat row, which is 52 households.

	Dog	**No Dog**	**Total**
Cat	52		180
No Cat			
Total	95		250

Since each row and column must sum to their totals, we can use subtraction to find the missing numbers as shown below.

	Dog	No Dog	Total
Cat	52	180-52 = **128**	180
No Cat	95-52 = **43**	155 - 128 = **27** or 70 – 43 = **27**	250 – 180 = **70**
Total	95	250 – 95 = **155**	250

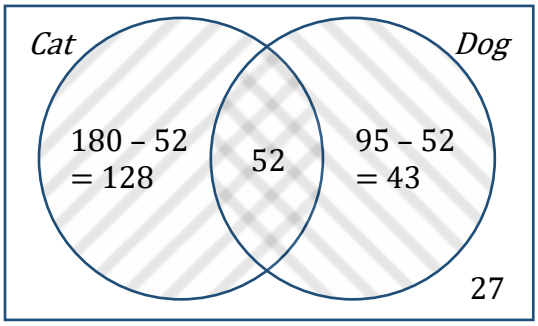
Now that we have our contingency table completed, notice that the numbers in the central four cells add to the grand total as shown in the table on the left. The total row and the total column also add to the grand total as shown in the right table.

	Dog	No Dog	Total
Cat	52	128	180
No Cat	43	27	70
Total	95	155	250

	Dog	No Dog	Total
Cat	52	128	180
No Cat	43	27	70
Total	95	155	250

Contingency Tables and Venn Diagrams

If the subtractions we just did seem familiar, they should! This is very similar to what we did for reporting data with a Venn diagram. The Venn diagram for this data is shown to the right.



We also subtracted the intersection from the total of the cat and dog owners to find numbers in the crescent regions.

Notice that the numbers in the four regions of the Venn diagram are the same as the four cells in the center of the contingency table and add to the grand total.

"And" Statements

Now we can use the contingency table or the Venn diagram to determine the percentage of households that meet certain conditions. For instance, what percent of those surveyed own a cat **and** do not own a dog? In the Venn diagram, this is 128 households in the cat only region.

In the contingency table we see the 128 households at the intersection of the row of households who own a cat and the column of households who do not own a dog. As a percentage, the total number of households surveyed, is

	Dog	No Dog	Total
Cat	52	128	180
No Cat	43	27	70
Total	95	155	250

$\frac{128}{250} = 0.512$ or 51.2% that have a cat and no dog.

"Or" Statements

How about the percentage of households surveyed that have a cat **or** a dog? We know from Venn diagrams that the inclusive *or* includes the number of households who own a cat only, a dog only, and both a cat and a dog, or $128 + 52 + 43 = 223$ households. As a percentage of the total surveyed, we get

	Dog	No Dog	Total
Cat	52	128	180
No Cat	43	27	70
Total	95	155	250

$\frac{223}{250} = 0.892$ or 89.2% of households in the sample have a dog **or** a cat (or both).

We can get the same answer from the contingency table. by adding the cells for households who have a cat and not a dog, a dog and not a cat, and the households that have both a cat and a dog. This also gives us 223 households.

There is another way to calculate an *or* statements from a contingency table. We could add the row and column totals for having a cat and having a dog, but then we have counted the 52 households in the intersection twice. We can subtract that number to get $180 + 95 - 52 = 223$ households with a dog or a cat, which we know is 89.2% of those surveyed.

Conditional Statements

Another question we can answer using a contingency table is what percentage of dog owning households also own a cat? In this case the group that we are interested in isn't every household surveyed (the grand total), but just those households that own a dog.

We call this a **conditional** statement because we are only considering the households with a certain condition. If we focus on the column representing the households that own a dog, we see that there is a total of 95 households with a dog, and that 52 of those 95 households also have a cat. Therefore,

	Dog	No Dog	Total
Cat	**52**	128	180
No Cat	43	27	70
Total	**95**	155	250

$\frac{52}{95} = 0.547$ or 54.7% of the households with a dog also have a cat. Another way to phrase this conditional statement is, "What percent of households have a cat **given** they have a dog." You will see the word given quite a bit in this chapter and that makes the denominator change. It is also possible to find this conditional percentage using the Venn diagram by taking the number in the intersection and dividing it by the total in the whole dog circle.

Contingency Tables with More Than Two Categories

When there are only two categories for each variable, like yes/no questions, Venn diagrams and contingency tables provide basically the same information and can be used interchangeably. A Venn diagram works well for yes/no variables since a subject is either inside the circle (has the characteristic) or outside the circle (does not have the characteristic). If we have more than two possibilities for any of the variables, though, we cannot use a Venn diagram. We can use a contingency table, though. Here is an example where one variable has four categories and the other has three categories.

Example 1: 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should (i) be allowed to keep their jobs and apply for US citizenship, (ii) be allowed to keep their jobs as temporary guest workers but not be allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. Not sure was also an option (iv). The results of the survey by political ideology are shown below[1]. Use the contingency table to answer the questions.

	Conservative	Moderate	Liberal	Total
(i) Apply for citizenship	57	120	101	278
(ii) Guest worker	121	113	28	262
(iii) Leave the country	179	126	45	350
(iv) Not sure	15	4	1	20
Total	372	363	175	910

a. What percent of the sampled Tampa, Fl voters identified themselves as conservatives? To answer this question, we find the conservative column and look to the bottom cell for the total number of conservative voters and divide that by the total number of voters surveyed. This gives us $\frac{372}{910} = 0.409$ or 41% of the Tampa, Fl voters who identify as conservative.

b. What percent of the sampled voters are in favor of the citizenship option? For this question we find the apply for citizenship row, look across to find the total, and divide this by the total number of voters surveyed. We get

[1] SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012. Example adapted from Open Intro: Advanced High School Statistics, by Diez et al, used under CC-BY-SA 3.0.

$\frac{278}{910} = 0.305$ or 31% of these voters are in favor of the citizenship option.

c. What percent of the sampled voters identify themselves as conservatives **and** are in favor of the citizenship option? For this question we are looking for the cell that is the intersection of those who identify as conservative and those who are in favor of the citizen option. This cell has 57 voters, so we divide that by the total number of voters. This gives us
$\frac{57}{910} = 0.063$ or 6% of these voters identify as conservatives and are in favor of the citizenship option.

d. What percent of the sampled voters identify themselves as liberal **or** are in favor of the leaving the country option? The *or* in this question is inclusive, so we need to determine the number of voters who identify as liberal, who are in favor of the leaving the country option, or both.

	Conservative	Moderate	Liberal	Total
(i) Apply for citizenship	57	120	101	278
(ii) Guest worker	121	113	28	262
(iii) Leave the country	179	126	45	350
(iv) Not sure	15	4	1	20
Total	372	363	175	910

In terms of the individual cells, the number of voters who have the specified characteristics is the sum $179+126+101+28+45+1=480$, which we can divide by the total number of voters surveyed to get the percent. So, we have
$\frac{480}{910} = 0.527$ or 53% of the voters identify as liberal or are in favor of the leave the country option.
Another way to calculate this is to add the total number who identify as liberal (175 voters) and the total number who are in favor of the leave the country option (350 voters), then subtract the double counted cell (45 voters) who are liberal and in favor of the leave the country option: $175+350-45=480$.

e. What percent of the sampled voters who identify as conservatives are also in favor of the citizenship option? What percent of moderate and liberal voters share this view? As we saw before, these are conditional statements. For the first part of this question, we want to focus just on those voters who identify as conservatives, and from among that group determine the percent in favor of the citizenship option. We calculate that
$\frac{57}{372} = 0.153$ or 15% of conservative voters are in favor of the citizenship option.

For the second part, we want to focus on just those voters who identify as moderate, and from among that group determine the percent in favor of the citizenship option. Then we have

$\frac{120}{363} = 0.331$ or 33% of moderate voters are in favor of the citizen option.

Finally, we want to focus on just those voters who identify as liberal, and from among that group determine the percent in favor of the citizenship option. We calculate

$\frac{101}{175} = 0.577$ or 58% of liberal voters are in favor of the citizenship option.

Looking at these three percentages, it is clear that support of the citizenship option **depends** on political ideology. If support of the citizenship option were the same across political ideologies, then we would say that favoring the citizenship option and political ideology were **independent** of each other.

Empirical Probability

If our sample is representative of the population, then we can also interpret a percentage we calculate from a contingency table as a **probability**, or the likelihood that something will happen. Since a contingency table is constructed from data collected through sampling or an experiment, we call it an **empirical** or **experimental** probability. This is different from a **theoretical** probability which we will look at in the next section.

Finding Empirical Probabilities with a Contingency Table

Suppose that 60% of students in our class have a summer birthday (June, July, or August). Now suppose everyone's name and birth month are written on slips of paper and thrown into a bag. If we pull a slip of paper out of the bag at random, what is the probability that the selected student has a summer birthday? If you think there should be a 60% chance, you are right! The relative frequency of the characteristic of interest will be equal to its empirical probability. To write this as a probability statement, it would look like

$P(\text{summer birthday}) = 60\%$

Probability is a function named *P*, and the function is applied to what follows in the parentheses. Let's look at another example where we write probability statements and find empirical probabilities.

Example 2: A survey of licensed drivers asked whether they had received a speeding ticket in the last year and whether their car is red. The results of the survey are shown in the contingency table to the right.

	Speeding Ticket	No Speeding Ticket	Total
Red Car	15	135	150
Not Red Car	45	470	515
Total	60	605	665

Find the probability that a randomly selected survey participant:

a. has a red car.

b. has had a speeding ticket in the last year.
c. has a red car and has not had a speeding ticket in the last year.
d. has a red car or has had a speeding ticket in the last year.
e. has had a speeding ticket in the last year given they have a red car.
f. who has received a speeding ticket in the last year also has a red car.
g. What do the answers to b and e suggest about the relationship between owning a red car and getting a speeding ticket?

Here are the solutions:

a. To find $P(\text{red car})$, we divide the number of participants who own a red car by the total number of people surveyed:

$$P(\text{red car}) = \frac{150}{665}$$
$$= 0.226 \text{ or } 22.6\%$$

	Speeding Ticket	No Speeding Ticket	Total
Red Car	15	135	150
Not Red Car	45	470	515
Total	60	605	665

b. To find $P(\text{speeding ticket})$, we divide the number of participants who got a speeding ticket in the last year by the total number of people surveyed:

$$P(\text{speeding ticket}) = \frac{60}{665}$$
$$= 0.090 \text{ or } 9\%$$

	Speeding Ticket	No Speeding Ticket	Total
Red Car	15	135	150
Not Red Car	45	470	515
Total	60	605	665

c. To find $P(\text{red and no ticket})$, we find the intersection of the red car category and the no ticket category and divide by the total number of participants:

$$P(\text{red and no ticket}) = \frac{135}{665}$$
$$= 0.203 \text{ or } 20.3\%$$

	Speeding Ticket	No Speeding Ticket	Total
Red Car	15	135	150
Not Red Car	45	470	515
Total	60	605	665

d. To find $P(\text{red or ticket})$, we need to add those who drive a red car and did not have a speeding ticket (just red), those who had a speeding ticket and do not drive a red car (just ticket) and those who drive a red car and had a speeding ticket (both), and divide by the total number of participants:

$$P(\text{red or ticket}) = \frac{135+45+15}{665} = \frac{195}{665} = 0.293 \text{ or } 29.3\%$$

	Speeding Ticket	No Speeding Ticket	Total
Red Car	15	135	150
Not Red Car	45	470	515
Total	60	605	665

Recall from our earlier discussion that we could also calculate the *or* probability as:

$$P(\text{red or ticket}) = P(\text{red}) + P(\text{speeding ticket}) - P(\text{red and speeding ticket})$$
$$= \frac{150}{665} + \frac{60}{665} - \frac{15}{665}$$
$$= \frac{195}{665}$$

which gives us the same answer as counting the individual cells.

e. The probability $P(\text{speeding ticket given red car})$ is a conditional probability as we have seen before since it is conditional on the given characteristic occurring. In this problem, the given characteristic is owning a red car, so we isolate our attention to just the row of 150 red car owners and see how many have had a speeding ticket in the last year. Looking at the table, we see that there were 15 red car owners who had a speeding ticket in the last year, so we calculate:

	Speeding Ticket	No Speeding Ticket	Total
Red Car	15	135	150
Not Red Car	45	470	515
Total	60	605	665

$$P(\text{speeding ticket given red car}) = \frac{15}{150} = 0.10 \text{ or } 10\%$$

f. This question is also asking for a conditional probability, $P(\text{red car given speeding ticket})$, but it is phrased more like we would say it. In this case the given characteristic is that the person has received a speeding ticket, so we will isolate our attention to just the speeding ticket column. Among the 60 people who had a speeding ticket in the last year, we see that 15 also drove a red car. Now we can calculate the probability:

	Speeding Ticket	No Speeding Ticket	Total
Red Car	15	135	150
Not Red Car	45	470	515
Total	60	605	665

$$P(\text{red car given speeding ticket}) = \frac{15}{60}$$
$$= 0.25 \text{ or } 25\%$$

Notice that compared with part e, when we change the conditional characteristic, we change the denominator of the fraction.

g. In part b, we determined that there was a 9% chance of randomly selecting a participant who had received a speeding ticket in the last year. However, in part e we found that there was a 25% chance of receiving a ticket in the last year if the person had a red car. This seems to suggest that there is a higher likelihood of getting a speeding ticket if you own a red car. This means that getting a speeding ticket is **dependent** on whether the person drives a red car, since that increases the probability of getting a ticket. We cannot say, however, whether driving a red car makes you speed or whether people who tend to drive faster buy red cars.

Conditional Probabilities

We have mentioned conditional probabilities, which we find by isolating our attention to the given row or column. Here is another example of finding conditional probabilities.

Example 3: A home pregnancy test was given to a sample of 93 women, and their pregnancy was then verified by a blood test. The contingency table below shows the home pregnancy test and whether or not they were actually pregnant as determined by the blood test. Find the probability that a randomly selected woman in the sample

a. was not pregnant given the home test was positive.
b. had a positive home pregnancy test given they were not pregnant.

	Positive Test	Negative Test	Total
Pregnant	70	4	74
Not Pregnant	5	14	19
Total	75	18	93

Here are the solutions:

a. Since we are given the home test result was positive, we are limited to the 75 women in the positive test column, of which 5 were not pregnant. This gives:

	Positive Test	Negative Test	Total
Pregnant	70	4	74
Not Pregnant	5	14	19
Total	75	18	93

$$P(\text{not pregnant given positive test}) = \frac{5}{75}$$
$$= 0.067 \text{ or } 6.7\%$$

b. Since we are given the woman is not pregnant, we are limited to the 19 women in the not pregnant row, of which 5 had a positive test. This gives:

	Positive Test	Negative Test	Total
Pregnant	70	4	74
Not Pregnant	5	14	19
Total	75	18	93

$$P(\text{positive test given not pregnant}) = \frac{5}{19}$$
$$= 0.263 \text{ or } 26.3\%$$

This result is referred to as a false positive: A positive test result when the woman is not actually pregnant.

In this section we have learned about empirical probability. In the next section we will discuss another kind of probability that you may be familiar with – theoretical probability.

Exercises 4.1

A professor gave a test to students in a morning class and the same test to the afternoon class. The grades are summarized below. Use the table for questions 1-2.

	A	B	C	Total
Morning Class	8	18	13	39
Afternoon Class	10	4	12	26
Total	18	22	25	65

1. If one student was chosen at random, find each probability:
   a. P(in the morning class)
   b. P(earned a C)
   c. P(earned an A and was in the afternoon class)
   d. P(earned an A given the student was in the morning class)

2. What is the probability that a student who earned a B was in the afternoon class? What's the probability that a student in the afternoon class earned a B? Explain the difference between these two quantities.

3. The contingency table below shows the number of credit cards owned by a group of individuals below the age of 35 and above the age of 35.

	Zero	One	Two or more	Total
Between the ages of 18-35	9	5	19	33
Over age 35	18	10	20	48
Total	27	15	39	81

If one person was chosen at random, find each probability:
   a. P(had no credit cards)
   b. P(had one credit card)
   c. P(had zero and is over 35)
   d. P(had zero credit cards given that the person under 35)

4. After reviewing the data, it was decided that more detail should be shown. The category of 35+ is divided into two groups and "Two or more" is turned into "Two" and "3 or more." Fill out the missing information.

	Zero	One	Two	Three or More	Total
Between the ages of 18-35	9	5		9	33
35-65	10	6	10		27
Over 65				8	
Total	27	15			81

5. Fill out the missing values in the contingency table below. This table will also be used for question 6.

	Heads	Tails	Total
Coin A	60		100
Coin B		2	2
Coin C	6,000		10,000
Total			

6. Using the table above to answer the following:
   a. Find the percentage of heads for coins A, B and C.
   b. If you knew one coin was weighted, which coin would you most suspect and why?

Section 4.2 Theoretical Probability

As we saw in the last section, the **probability** of a specified event is the chance or likelihood that it will occur. We calculated empirical or experimental probabilities using contingency tables. In this section, we will focus on **theoretical** probability and compare the two types.

Basic Probability Concepts

Let's begin with a brief introduction to some of the language and basic concepts of theoretical probability.

Experiment

If you roll a die, pick a card from a deck of playing cards, or randomly select a person and observe their hair color, you are conducting an **experiment**.

Events and Outcomes

The result of an experiment is an **outcome**, and a particular outcome, like rolling a five on a die, is called an **event.** An event can be a **simple event** or combination of outcomes, called a **compound event**.

Sample Space

The **sample space** is the set of all possible outcomes. For example, if we roll a six-sided die, the sample space S is the set $S = \{1, 2, 3, 4, 5, 6\}$.

> Example 1: If we roll an eight-sided die, describe the sample space and give at least two examples of simple events and compound events.
>
> The sample space is the set of all possible outcomes, or equivalently, all simple events: $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$
>
> Examples of simple events are rolling a 1, rolling a 5, rolling a 6, and so on. Examples of compound events include rolling an even number, rolling a 5 or a 3, and rolling a number that is at least 4.

Equally Likely Outcomes

When the outcomes of an experiment are equally likely, we can calculate the probability of an event as the number of ways it can happen out of the total number of outcomes.

Theoretical Probability

$$P(E) = \frac{\text{number of outcomes corresponding to the event E}}{\text{total number of equally-likely outcomes}}$$

We can write the result as a simplified fraction or as a decimal or percent.

Example 2: Write the sample space for the sum of two six-sided dice and determine whether the outcomes are equally likely.

The sample space for the sum of two six-sided dice is $S = \{2,3,4,5,6,7,8,9,10,11,12\}$. The different sums, however, are not equally likely. If we look at a table of the different possible outcomes when rolling two dice, we see that there are 36 possible combinations. We will summarize this in a table by listing each outcome in the sample space. Then we find the probabilities by counting the number of ways each sum can occur and dividing it by 36.

	1	2	3	4	5	6
1	1+1 = 2	1+2 = 3	1+3 = 4	1+4 = 5	1+5 = 6	1+6 = 7
2	2+1 = 3	2+2 = 4	2+3 = 5	2+4 = 6	2+5 = 7	2+6 = 8
3	3+1 = 4	3+2 = 5	3+3 = 6	3+4 = 7	3+5 = 8	3+6 = 9
4	4+1 = 5	4+2 = 6	4+3 = 7	4+4 = 8	4+5 = 9	4+6 = 10
5	5+1 = 6	5+2 = 7	5+3 = 8	5+4 = 9	5+5 = 10	5+6 = 11
6	6+1 = 7	6+2 = 8	6+3 = 9	6+4 = 10	6+5 = 11	6+6 = 12

Sum	Probability
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

From the probability table we can see that rolling a sum of 7 has the highest probability and rolling a 2 or a 12 have the lowest probabilities.

Example 3: Suppose we roll a fair six-sided die. Calculate the probability of:

a. rolling a 6.
b. rolling a number that is at least 4.
c. rolling an even number.
d. rolling a 5 or a 3.

Recall that the sample space is $S = \{1,2,3,4,5,6\}$. Since each of the outcomes in the sample space is equally likely, we can find the probability of each event by counting the number of outcomes corresponding to the event and dividing by 6, the total number of equally likely outcomes.

a. There is only one way to roll a 6, so

$P(\text{rolling a } 6) = \frac{1}{6}$ or 16.7%

There is a 16.7% chance of rolling a 6.

b. In probability we will often come across the phrases "at least" and "at most." **At least** means that value or greater. **At most** means that value or less. Since we are looking for the probability of rolling a number that is at least 4, we need the number of outcomes that are 4 or greater. There are 3 values that meet this condition: 4, 5, and 6. The probability is

$$P(\text{Rolling a number that is at least 4}) = \frac{3}{6} = \frac{1}{2} \text{ or } 50\%$$

c. Half of the numbers on a die are even, so we calculate:

$$P(\text{rolling an even number}) = \frac{3}{6} = \frac{1}{2} \text{ or } 50\%$$

d. There are two ways to roll a 5 or a 3, so

$$P(\text{Rolling a 5 or a 3}) = \frac{2}{6} = \frac{1}{3} \text{ or } 33.3\%$$

There is a 33.3% chance of rolling a 5 or a 3.

Example 4: Suppose you have a bag containing 14 sweet cherries and 6 sour cherries. If you pick a cherry at random, what is the probability it will be sweet?

Each of the cherries are equally likely to be selected since our selection is random and we can assume there is no way to distinguish one cherry from another. This means that the probability of selecting a sweet cherry will be equal to the number of sweet cherries in the bag dived by the total number of cherries in the bag. Since there are 14 sweet cherries and a total of 20 cherries in the bag, we have:

$$P(\text{sweet}) = \frac{14}{20} = \frac{7}{10} \text{ or } 70\%$$

There is a 70% chance of selecting a sweet cherry from the bag.

Certain and Impossible Events

A probability is always a value between 0 and 1, or from 0% to 100%. If the probability of an event is 0 there are no outcomes that correspond with that event and we say it is **impossible**. If the probability of an event is 1 then every outcome corresponds to that event and we say it is **certain**.

Example 5:

a. What is the probability of rolling an odd or even number on a six-sided die? Since all the numbers are either even or odd, this event includes all of the outcomes in the sample space. This event is certain.

$$P(\text{odd or even}) = \frac{6}{6} = 1.$$

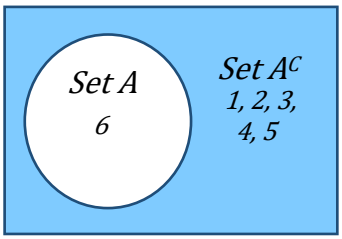
b. What is the probability of rolling an 8 on a six-sided die? Since 8 is not one of the outcomes in the sample space, the event is impossible.

$$P(\text{roll an 8}) = \frac{0}{6} = 0.$$

Complementary Events

Just as we saw in the logic chapter, the **complement** of an event A means A does not happen. We can refer to the complement as ***not A*** or A^C. For example, consider the experiment of rolling a six-sided die and the simple event A = rolling a 6. The complement of event A is everything in the sample space that is not a 6: $A^C = \{1, 2, 3, 4, 5\}$. Recall that we can illustrate the complement using a Venn diagram as shown below.

Notice that the outcomes from set A and the outcomes from set A^C will together equal the universal set, which is the sample space in probability. The probabilities must add up to 1 or 100%. Therefore, we can use subtraction to find the probability of a complement.



Complement of an Event

$$P\left(A^C\right) = 1 - P(A)$$

Example 6: If you roll an eight-sided die, what's the probability you don't get a 6?

Not rolling a 6 is the complement of rolling a 6, which is easier to calculate. Since there are 8 possible numbers to roll, we have:

$$P(\text{not rolling a 6}) = 1 - P(\text{rolling a 6})$$

$$= 1 - \frac{1}{8}$$

$$= \frac{7}{8} \text{ or } 0.875$$

There is an 87.5% chance of not rolling a 6.

Experimental vs. Theoretical Probability

Now that we have calculated experimental and theoretical probabilities, we can compare them. When we flip a coin, we say there is a 50% chance of getting heads. This is a theoretical probability because there are two equally likely outcomes – heads and tails – so we expect to get heads half of the time. But if you flip a coin, say 100 times, will you get heads exactly 50 times? Maybe, but you are more likely to get some number around 50 times. The number of heads you actually observe out of the total number of times you flip the coin is the experimental probability.

Example 7: The table shows the numbers that came up after rolling a six-sided die 10 times. What is the experimental probability of rolling a 6? What is the theoretical probability of rolling a 6?

Roll	1	2	3	4	5	6	7	8	9	10
Outcome	3	1	4	6	6	6	1	3	5	1

To find the experimental probability of rolling a 6, it would be helpful to change this into a frequency table. We list all the possible outcomes and count how many times each occurred. According to our frequency table, we see that a 6 was rolled three times, so the experimental probability of rolling a 6 is

Outcome	Frequency
1	3
2	0
3	2
4	1
5	1
6	3

$$P(\text{roll } 6) = \frac{3}{10} \text{ or } 30\%.$$

Theoretically, however, we would expect the number 6 to come up 1 out of 6 times since there are 6 equally likely outcomes. Thus, the theoretical probability of rolling a 6 is

$$P(\text{roll } 6) = \frac{1}{6} \text{ or } 16.7\%$$

The Law of Large Numbers

As we saw in the previous example, theoretical and experimental probabilities are not necessarily equal. However, experimental probability will eventually approach theoretical probability as we conduct more and more trials. This phenomenon is called the **Law of Large Numbers**. This means if you flip a coin a small number of times, the experimental probability is likely to be different each time and could be very different from the theoretical probability. But if you flip a coin a large number of times, the experimental probability becomes very close to the theoretical probability of 50%.

The Law of Large Numbers is extremely powerful in that it allows us to approximate the theoretical probability of complex events – like changes in beliefs and opinions, likelihood of natural disasters, climate change effects – through repeated sampling and simulation.

Probability of Compound Events

Now that we have the basics in place, let's look at some compound probability problems that we will be studying in this course.

"And" Probabilities

As we saw with truth tables, the event ***A and B*** refers to an event where both *A* and *B* occur. These events may occur at the same time or they could happen in a sequence such as ***A and then B***. How we calculate the theoretical probability of the event *A and B* (or *A and then B*) depends on whether the two events are independent or dependent.

Independent and Dependent Events

Two events *A* and *B* are **independent** if the probability of *B* occurring is the same whether or not *A* occurs. If the probability of *B* is affected by the occurrence of *A*, then we say that the events are **dependent**.

Coin flips and die rolls are common examples of independent events – flipping heads does not change the probability of flipping heads the next time, nor does rolling a six change the probability that the next roll will be a six.

Another type of event is a selection event, such as randomly selecting or drawing items from a bag, etc. These are also independent if we **draw with replacement**. By replacing the item, we reset the probability back to what it was before we made the selection. Since the probability of each selections is the same as the first selection, the events are independent.

If we draw **without replacement**, however, like selecting multiple people for a committee, we change the total number of possible outcomes, thereby changing the probability of subsequent selections. Therefore, if we draw without replacement, the events will be **dependent**.

Example 8: Determine whether the following events are independent or dependent.

a. Flipping a coin twice and getting heads both times.
b. Selecting a president and then a vice president at random from a pool of five equally qualified individuals.
c. The event that it will rain in Portland tomorrow and the event that it will rain in Beaverton tomorrow.
d. Wearing your lucky socks and getting an A on your exam.

Here are the answers:

a. The probability of getting heads on the first flip is 0.5 or 50%. After flipping heads, the probability of getting heads on the second flip is still 0.5 or 50%. Since the probability of flipping heads on the second flip did not change because we flipped heads on the first flip, the events are **independent**.
b. Since two different people will be put in the role of president and vice president, we are drawing without replacement and the events are therefore **dependent**.
c. If it is raining in Portland it is more likely that it will rain in Beaverton, so the events are **dependent**.
d. Although there may some sort of placebo effect at play in terms of confidence and persistence, the socks you wear do not have a direct effect on how well you do on your exam, so these events are **independent**.

To calculate *and* probabilities we multiply, but we need to determine whether the events are independent or dependent. If they are independent, then we can multiply the individual probability of each event because one does not affect the other. If the events are dependent, then we need to multiply by the conditional probability based on what has previously happened. Here is a summary of this.

***"And"* Probabilities**

If events A and B are independent, then $P(A \text{ and } B) = P(A) \cdot P(B)$

If events A and B are dependent, then $P(A \text{ and } B) = P(A) \cdot P(B \text{ given } A)$

The probability of *B given A* is called a **conditional probability** since it depends, or is conditional, on *A* occurring. We saw examples of conditional probability when we looked at contingency tables in the previous section.

Example 9: Suppose you have a bag containing 6 red Legos, 4 green Legos, and 3 black Legos. What is the probability of selecting

a. two red Legos in a row if we put the first red Lego back in the bag?
b. two red Legos in a row if we don't put the first Lego back in the bag?
c. a red Lego and then a green Lego if we do not put the red Lego back in the bag?

Here are the solutions:

a. Since the outcomes are equally likely, the probability of selecting a red Lego is the number of red Legos divided by the total number of Legos, or $P(\text{red}) = \frac{6}{13}$.

   If we replace the red Lego we selected (selections are independent), we go back to having 6 red Legos in the bag of 13 Legos total. Therefore,

$$P(\text{red and then red}) = P(\text{red}) \cdot P(\text{red})$$
$$= \frac{6}{13} \cdot \frac{6}{13}$$
$$= 0.213 \text{ or } 21.3\%$$

b. If we do not replace the first red Lego (selections are dependent), then on our second draw there will only be 5 red Legos remaining, and 12 Legos in total. Therefore,

$$P(\text{red and then red}) = P(\text{red}) \cdot P(\text{red given red taken out})$$
$$= \frac{6}{13} \cdot \frac{5}{12}$$
$$= 0.192 \text{ or } 19.2\%$$

c. The probability of selecting a red Lego on the first draw is the same as in parts a and b. Since we are not putting the red Lego back into the bag, we will have only 12 Lego left in total, of which 4 are green. Therefore,

$$P(\text{red and then green}) = P(\text{red}) \cdot P(\text{green given red taken out})$$
$$= \frac{6}{13} \cdot \frac{4}{12}$$
$$= \frac{6}{13} \cdot \frac{1}{3}$$
$$= 0.154 \text{ or } 15.4\%$$

Let's look at an example where we repeat an event many times.

Example 10: Suppose there is a 6% chance you will receive a citation if you ride the MAX train without a ticket. What is the probability that you get away without a single citation if you ride without purchasing a ticket for 20 days this month?

The first thing we want to recognize is that this question is essentially asking for the probability of no citation and no citation and no citation…. twenty times (one for each day you ride without buying a ticket). Since the outcomes are connected by an "and", we know we will be multiplying the probabilities. In this case the the outcomes are independent (you are not more or less likely to get a citation if you already received a citation). Therefore,

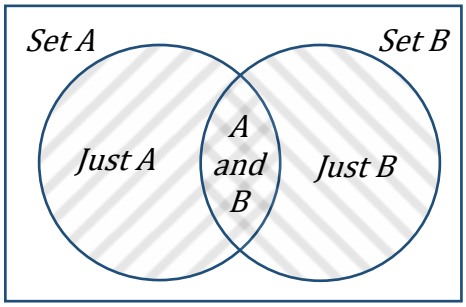
$$\begin{aligned} P(\text{no citation in 20 rides}) &= P(\text{no citation on a single ride})^{20} \\ &= (1-0.06)^{20} \\ &= (0.96)^{20} \\ &= 0.442 \text{ or } 44.2\% \end{aligned}$$

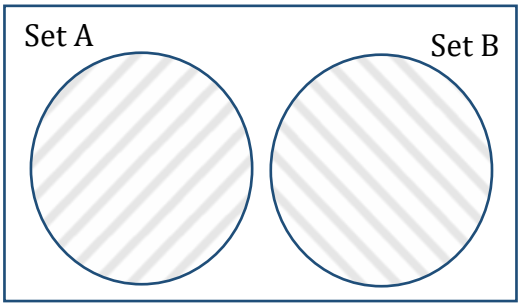
"Or" Probabilities

The event ***A or B*** refers to an event that includes the outcomes of *A* or *B* or *both*. We have seen the inclusive *or* both in terms of sets and logic, and in terms of contingency tables. The way we calculate the probability of *A or B* depends on whether the events have characteristics that are overlapping or disjoint.

Overlapping or Disjoint Sets

Recall that **disjoint** means the same thing as **not overlapping**. Just like we saw in the logic and sets chapter, the set diagram on the left shows overlapping sets and the set diagram on the right shows disjoint sets.





To apply this to probability, we will look at an example of events that have overlapping characteristics, such as color and shape.

Example 11: A prize machine is filled with 10 yellow erasers, 6 green erasers, 4 red pencil sharpeners, 8 yellow pencil sharpeners, and 5 red bouncy balls. Each prize is inside a plastic sphere, and the spheres are well mixed in the prize machine. Each game will get you just one prize. Determine the probability of

a. getting a yellow prize.
b. getting a red or yellow prize.
c. getting a prize that is yellow or an eraser.

Here are the answers:

a. Since yellow is a single event, we just need to know how many prizes there are in total, and how many of the prizes are yellow. The yellow prizes include the 10 yellow erasers and the 8 yellow pencil sharpeners.
$P(\text{yellow}) = \frac{18}{33}$.

b. For a red or yellow prize, the set of red and the set of yellow do not overlap. They are disjoint sets, so we will add the probability of getting a red prize to the probability of getting a yellow prize.
$$P(\text{red or yellow}) = P(\text{red}) + P(\text{yellow})$$
$$= \frac{9}{33} + \frac{18}{33}$$
$$= \frac{27}{33}$$

c. To find the probability of getting a prize that is yellow or an eraser, we need to be careful because these are overlapping sets. There are two ways to calculate this, and it is a lot like what we did with contingency tables.
The first way is to add all the items separately, being careful not to double count.
$$P(\text{yellow or eraser}) = P(\text{yellow eraser}) + P(\text{yellow pencil sharpener}) + P(\text{green eraser})$$
$$= \frac{10}{33} + \frac{8}{33} + \frac{6}{33}$$
$$= \frac{24}{33}$$
The second way is to count the total of yellow items and the total of erasers, but the yellow erasers are in both sets, or the overlap. We would be counting them twice and so we subtract their probability.
$$P(\text{yellow or eraser}) = P(\text{yellow}) + P(\text{eraser}) - P(\text{yellow and eraser})$$
$$= \frac{18}{33} + \frac{16}{33} - \frac{10}{33}$$
$$= \frac{24}{33}$$

Here is a summary of how we found the *or* probabilities.

"Or" Probabilities

If the sets are disjoint, $P(A \text{ or } B) = P(A) + P(B)$

If the sets are overlapping, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

We could also use the overlapping formula as a general formula, because in the case of disjoint sets, there is no intersection and $P(\text{A and B}) = 0$. Here is another example with overlapping events.

Example 12: What is the probability of rolling two dice and getting a pair or a sum of 6?

For complicated events it's a good idea to list all of the outcomes. Looking at the table of outcomes, we see that there are 6 outcomes that are pairs out of the 36 possible outcomes, and there are 5 outcomes that add to 6. We can also see that there is one outcome that is both a pair and a sum of 6, so the events are overlapping.

	1	**2**	**3**	**4**	**5**	**6**
1	1+1 = 2	1+2 = 3	1+3 = 4	1+4 = 5	1+5 = 6	1+6 = 7
2	2+1 = 3	2+2 = 4	2+3 = 5	2+4 = 6	2+5 = 7	2+6 = 8
3	3+1 = 4	3+2 = 5	3+3 = 6	3+4 = 7	3+5 = 8	3+6 = 9
4	4+1 = 5	4+2 = 6	4+3 = 7	4+4 = 8	4+5 = 9	4+6 = 10
5	5+1 = 6	5+2 = 7	5+3 = 8	5+4 = 9	5+5 = 10	5+6 = 11
6	6+1 = 7	6+2 = 8	6+3 = 9	6+4 = 10	6+5 = 11	6+6 = 12

As in the last example, there are two ways to do this.

If we add all of the shaded squares without double counting, we get:

$$P(\text{pair or sum of 6}) = P(\text{pair}) + P(\text{sum of 6 that haven't been counted})$$
$$= \frac{6}{36} + \frac{4}{36}$$
$$= \frac{10}{36}$$
$$= \frac{5}{18}$$

To use the subtraction method, we need to add the probability of rolling a pair to the probability of rolling a sum of 6 and subtract the overlap. Thus we have:

$$P(\text{pair or sum of 6}) = P(\text{pair}) + P(\text{sum of 6}) - P(\text{pair and a sum of 6})$$
$$= \frac{6}{36} + \frac{5}{36} - \frac{1}{36}$$
$$= \frac{10}{36}$$
$$= \frac{5}{18}$$

Now that we have looked at empirical and theoretical probability, we will be able to use them for something very important in the next section – expected value.

Exercises 4.2

1. A ball is drawn randomly from a jar containing 6 red marbles, 2 white marbles, and 5 yellow marbles. Find the probability of:
    a. Drawing a white marble.
    b. Drawing a red marble.
    c. Drawing a green marble.
    d. Drawing two yellow marbles if you draw with replacement.
    e. Drawing first a red marble then a white marble if marbles are drawn without replacement.

2. Compute the probability of tossing a six-sided die and getting
    a. an even number.
    b. a number less than 3.

3. Compute the probability of rolling a 12-sided die and getting
    a. a number other than 8.
    b. a 2 or 7.

4. A six-sided die is rolled twice. What is the probability of getting
    a. a 6 on both rolls?
    b. a 5 on the first roll and an even number on the second roll?

5. Suppose that 21% of people own dogs. If you pick two people at random, what is the probability that neither own a dog?

6. At some random moment, you look at your clock and note the minutes reading.
    a. What is probability the minutes reading is 15?
    b. What is the probability the minutes reading is 15 or less?

7. What is the probability of flipping a coin three times
    a. and getting a head each time?
    b. not getting a head at all?

8. What is the probability of rolling two six-sided dice
    a. and getting a sum greater than or equal to 7?
    b. getting an even sum or a sum greater than 7?

9. A box contains four black pieces of cloth, two striped pieces, and six dotted pieces. A piece is selected randomly and then placed back in the box. A second piece is selected randomly. What is the probability that
    a. both pieces are dotted?
    b. the first piece is black, and the second piece is dotted?
    c. one piece is black, and one piece is striped?

Section 4.3 Expected Value

Expected value is one of the useful probability concept we will discuss. It has many applications, from insurance policies to making financial decisions, and it's one thing that the casinos and government agencies that run gambling operations and lotteries may hope most people never learn about.

Expected Value

The **expected value** is the average gain or loss of an event if the procedure is repeated many times. To help get a better understanding of what expected value is and how it is used, consider the following scenario:

> You are commissioned to design a game for a local carnival. Your proposed game will have players roll a six-sided die. If it comes up 6, they win $10. If not, they get to roll again. If they get a 6 on the second roll, then they win $3. If they do not get a 6 on the second roll, they lose. With the game design complete, you now need to decide how much the carnival game owner should charge players in order to make a profit over the long run.

To make a profit, the carnival needs to know how much they will pay in winnings, on average, over the long run and charge more than that. In other words, they must charge more than the expected value of the game.

One way to organize the outcomes and probabilities is with a probability model. A **probability model** or **probability distribution** is a table listing the possible outcomes and their corresponding probabilities. The outcomes will be the amounts a player can win, and we will calculate the probabilities using what we have learned about theoretical probability.

As we have seen with complements, probabilities in a probability distribution must add to 1, so that is important to check. Here is the probability model for the carnival game:

Outcome ($ won)	Rolling Event	Probability
$10	Roll a 6 on the first roll	$P(\text{roll a } 6) = \frac{1}{6}$
$3	Roll not a 6 on the first roll and a 6 on the second roll	$P(\text{roll not a } 6 \text{ then roll a } 6) = \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{36}$
$0	Roll not a 6 on the first roll and not a 6 on the second roll	$P(\text{roll not a } 6 \text{ then not a } 6) = \frac{5}{6} \cdot \frac{5}{6} = \frac{25}{36}$

Think of the expected value as a weighted average. We could take the average of $10, $3, and $0, but they are not all equally likely. It is much more likely to win $0 than to win $10. So, to find the average, we multiply each outcome by the chance it will happen and add the products together.

Expected Value

Multiply each outcome by its probability and add up the products

In this case we have:

$$\text{Expected Winnings} = \$10\left(\frac{1}{6}\right)+\$3\left(\frac{5}{36}\right)+\$0\left(\frac{25}{36}\right)$$
$$= \$2.08$$

This tells us that over the long run, players can expect to win $2.08 per game. This also means that the carnival owner will be paying out an average of $2.08 per game! Since the carnival owner would rather not lose money by paying players over the long run we need to make sure to charge players enough to offset the average payout.

If the carnival owner charges exactly $2.08 to play, the game is considered a **fair game** since the expected winnings would be $0. In a fair game, the player isn't expected to win anything, nor is the owner expected to earn anything over the long run. However, if the carnival owner charges the player more than $2.08 to play, they will earn money over the long run.

Suppose you suggest charging $5 to play. We can determine the **net** winnings by subtracting the $5 the player has to pay from their expected winnings. This gives us:

$$\text{Net player winnings} = \$2.08 - \$5.00$$
$$= \$ - 2.92$$

This means that over the long run, players can expect to lose an average of $2.92 each game they play, and the carnival owner can expect to earn an average of $2.92 per game over the long run. Here's another example.

Example 1: Pick4 is a game by the Oregon Lottery that costs $1 to play. In this game you pick 4 numbers in a specific pattern. If you get the exact sequence, you can in theory earn a lot of money. Suppose that the payouts are as follows. Determine the player's expected net winnings.

This table is not quite a complete probability distribution since it is missing one important outcome: when the player loses. In that case the prize is $0. We need to add a line for this. The prize for this missing outcome is $0, and since losing is the complement to winning *something*, the probability will be:

Prize ($)	Probability
$250	1/417
$500	1/1833
$1,000	1/1667
$1,500	1/2500

$$P(\text{win \$0}) = 1 - \left(\frac{1}{417} + \frac{1}{1833} + \frac{1}{1667} + \frac{1}{2500} \right)$$
$$= 1 - 0.0039$$
$$= 0.9961$$

Adding this information to the table gives a complete probability distribution. Now we can see that players are going to lose more than 99% of the time, so the expected value will be heavily weighted toward winning \$0.

Prize (\$)	Probability
\$250	1/417
\$500	1/1833
\$1,000	1/1667
\$1,500	1/2500
\$0	0.9961

$$\text{Expected Winnings} = 250\left(\frac{1}{417}\right) + 500\left(\frac{1}{1833}\right) + 1000\left(\frac{1}{1667}\right) + 1500\left(\frac{1}{2500}\right) + 0(0.9961)$$
$$= \$2.07$$

Therefore, the player's expected winnings are \$2.07, on average, over the long run. To find the expected **net** winnings, we subtract the cost to play. Since it costs \$1 to play,

$$\text{Net Expected Winings} = \$2.07 - \$1.00$$
$$= \$1.07$$

Assuming the given payouts are correct, this would be one game you would want to play for investment purposes since you can expect to earn \$1.07 per game, on average, over the long run. Play a million times, and you just might become a millionaire!

In general, if the expected value of a game is negative, it is not a good idea to play, since in the long run you will lose money. It would be better to play a game with a positive expected value (good luck trying to find one!), although keep in mind that even if the average winnings are positive it could be the case that most people lose money and one very fortunate individual wins a great deal of money.

Not surprisingly, the expected value for casino games is always negative for the player, and therefore positive for the casino. It must be positive for the casino, or they would go out of business! Players just need to keep in mind that when they play a game repeatedly, they should expect to lose money. 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.

Expected value is not only used to determine the average amount won and lost at casinos and carnivals, it also has applications in business and insurance, just to name a few. Let's look at a couple of those applications.

Example 2: For 3 months, a coffee shop tracked their morning sales of coffee, between 6am and 10am. The following results were recorded:

Number of cups sold	145	150	155	160	170
Probability	0.15	0.22	0.37	0.19	0.07

How many cups of coffee should they expect to sell each morning?

In this case the table tells us that 15% of the time they sell 145 cups of coffee between 6am and 10am, 22% of the time they sell 150 cups, 37% of the time they sell 155 cups, 19% of the time they sell 160 cups, and 7% of the time they sell 170 cups. Since the highest probability is associated with 155 cups, the expected value should lie somewhat close to this.

To find the expected number of coffees sold, we multiply each number of cups of coffee by its respective probability and then add the products.

$$\text{Expected Number of Coffees Sold} = 145(0.15)+150(0.22)+155(0.37)+160(0.19)+170(0.07)$$
$$= 154.4 \text{ cups of coffee}$$

This means that over the long run, the coffee shop can expect, on average, to sell around 154 cups of coffee each morning. This is an important tool for businesses since it helps inform them how much stock they should keep on hand.

Example 3: On average, a 40-year-old man in the US has a 0.242% chance of dying in the next year[2]. An insurance company charges $275 annually for a life insurance policy that pays a $100,000 death benefit. What is the expected value for the insurance company on this policy?

The first thing we want to do is organize the probabilities and outcomes in a probability distribution table. There are two outcomes – either the policy holder dies, and the insurance company pays the benefit, or the policy holder does not die, and they do not pay anything.

The probability of paying the death benefit is equal to the chance of the person dying in the next year, and the probability of paying nothing is equal to the complement of the chance of dying in the next year.

Insurance Payout	**Probability**
$100,000	0.00242
$0	1 – 0.00242 = 0.99758

[2] According to the estimator at http://www.numericalexample.com/index.php?view=article&id=91

Then we can calculate:

$$\begin{aligned}\text{Expected Payout} &= \$100{,}000(0.00242) + \$0(0.99758) \\ &= \$242\end{aligned}$$

So, the expected payout for the insurance company is $242, but they are charging $275 for the policy. Their net revenue would be

$$\begin{aligned}\text{Net Value to Insurance Company} &= \$275 - \$242 \\ &= \$33\end{aligned}$$

The insurance company is making, on average, $33 per policy per year. It shouldn't be too surprising because there are overhead costs and the insurance company can only afford to offer policies if they, on average, make money on them. But how much money should they make? As a consumer it is important to know about expected value.

Exercises 4.3

1. A bag contains 3 gold marbles, 6 silver marbles, and 28 black marbles. Someone offers to play this game: You randomly select on marble from the bag. If it is gold, you win $3. If it is silver, you win $2. If it is black, you lose $1.
    a. Make a probability model for this game.
    b. What is your expected value if you play this game?

2. A friend devises a game that is played by rolling a single six-sided die once. If you roll a 6, he pays you $3; if you roll a 5, he pays you nothing; if you roll a number less than 5, you pay him $1.
    a. Make a probability model for this game.
    b. Compute the expected value for this game.
    c. Should you play this game?

3. A company wants to offer a 2-year extended warranty in case their product fails after the original warranty period but within 2 years of the purchase. They estimate that 0.7% of their products will fail during that time, and it will cost them $350 to replace a failed product. If they charge $48 for the extended warranty, what is the company's expected profit or loss on each warranty sold?

4. An insurance company estimates the probability of an earthquake in the next year to be 0.0013. The average damage done by an earthquake it estimates to be $60,000. If the company offers earthquake insurance for $100, what is their expected value of the policy?

5. You purchase a raffle ticket to help out a charity. The raffle ticket costs $5. The charity is selling 2000 tickets. One of them will be drawn and the person holding the ticket will be given a prize worth $4000. Compute the expected value for this raffle.

6. At the local fair there is a game in which folks are betting where a chicken will poop on a 5 by 5-foot grid. (There are 25, 1 by 1 squares to choose from) You can buy a 1 by 1-foot square for $10 and if the chicken poops on your square you win $100. Find the expected value for this game.

7. Create a problem using the concept of expected value. Possible topics include insurance policies, financial decisions, gambling and lotteries. Determine the expected value of the situation you created.