IDEAS OF MATHEMATICS



Patrick Kenyon
Mt. San Jacinto College



Ideas of Mathematics

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

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

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



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

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

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

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

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

This text was compiled on 10/17/2026

TABLE OF CONTENTS

Licensing

1: Number Sense

1.1: Numbers and Operations

1.2: Percent

1.3: Algebra Essentials

1.4: Graphing

2: Finance

2.1: Budgeting

2.2: Simple Interest

2.3: Compound Interest

2.4: Installments

3: Set Theory and Logic

3.1: The Basics

3.2: Operations with Sets

3.3: Venn Diagrams

3.4: Inductive and Deductive Reasoning

3.5: Using Venn Diagrams to Analyze Arguments

3.6: Common Valid and Invalid Arguments

3.7: Truth Tables

3.8: Analyzing Arguments with Truth Tables

4: Descriptive Statistics

4.1: Introduction to Statistics

4.2: Collecting Data

4.3: Visual Summaries of Data

4.4: Other Summaries

4.5: Measures of the Center and Variation of the Data

4.6: Measures of the Location of the Data and Outliers

5: Probability

5.1: Introduction on Probability

5.2: Expected Value

5.3: Probability Rules- “And” and “Or”

5.4: Counting Methods

5.5: Conditional Probabilities

6: Inferential Statistics

6.1: Discrete Random Variables

6.2: Continuous Random Variables

6.3: Normal Distributions
6.4: Estimating Population Mean
6.5: Estimating Population Proportion

7: Modeling

7.1: Introduction to Modeling
7.2: Modeling with Linear Functions
7.3: Fitting Linear Models to Data
7.4: Modeling with Exponential Functions
7.5: Fitting Exponential Models to Data
7.6: Putting It All Together
7.7: Modeling with Quadratic Functions
7.8: Scatter Plots

8: Additional Topics

Index

Glossary

Detailed Licensing

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

Licensing

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

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

1: NUMBER SENSE



- 

1.1: Numbers and Operations

1.2: Percent

1.3: Algebra Essentials

1.4: Graphing

This page titled 1: Number Sense is shared under a license and was authored, remixed, and/or curated by LibreTexts.

1.1: Numbers and Operations

Learning Objectives

- Classify a real number as a natural, whole, integer, rational, or irrational number.
- Perform calculations using order of operations.
- Round decimals

It is often said that mathematics is the language of science. If this is true, then the language of mathematics is numbers. The earliest use of numbers occurred 100 centuries ago in the Middle East to count, or enumerate items. Farmers, cattlemen, and tradesmen used tokens, stones, or markers to signify a single quantity—a sheaf of grain, a head of livestock, or a fixed length of cloth, for example. Doing so made commerce possible, leading to improved communications and the spread of civilization.

Three to four thousand years ago, Egyptians introduced fractions. They first used them to show reciprocals. Later, they used them to represent the amount when a quantity was divided into equal parts.

But what if there were no cattle to trade or an entire crop of grain was lost in a flood? How could someone indicate the existence of nothing? From earliest times, people had thought of a “base state” while counting and used various symbols to represent this null condition. However, it was not until about the fifth century A.D. in India that zero was added to the number system and used as a numeral in calculations.

Clearly, there was also a need for numbers to represent loss or debt. In India, in the seventh century A.D., negative numbers were used as solutions to mathematical equations and commercial debts. The opposites of the counting numbers expanded the number system even further.

Because of the evolution of the number system, we can now perform complex calculations using these and other categories of real numbers. In this section, we will explore sets of numbers, calculations with different kinds of numbers, and the use of numbers in expressions.

Classifying a Real Number

The numbers we use for counting, or enumerating items, are the **natural numbers**: $1, 2, 3, 4, 5$ and so on. We describe them in set notation as $\{1, 2, 3, ...\}$ where the ellipsis $(\cdots)$ indicates that the numbers continue to follow the pattern. The natural numbers are, of course, also called the **counting numbers**. Any time we enumerate the members of a team, count the coins in a collection, or tally the trees in a grove, we are using the set of natural numbers. The set of **whole numbers** is the set of natural numbers plus zero: $\{0, 1, 2, 3, ...\}$.

The set of **integers** adds the opposites of the natural numbers to the set of whole numbers: $\{\cdots, -3, -2, -1, 0, 1, 2, 3, \cdots\}$.It is useful to note that the set of integers is made up of three distinct subsets: negative integers, zero, and positive integers. In this sense, the positive integers are just the natural numbers. Another way to think about it is that the natural numbers are a subset of the integers.

$$\begin{array}{ccc} \text{negative integers} & \text{zero} & \text{positive integers} \\ \cdots, -3, -2, -1 & 0 & 1, 2, 3, \cdots \end{array} \tag{1.1.1}$$

The set of **rational numbers** is written as $\{\frac{m}{n} | \text{m and n are integers and } n \neq 0\}$. Notice from the definition that rational numbers are fractions (or quotients) containing integers in both the numerator and the denominator, and the denominator is never 0. We can also see that every natural number, whole number, and integer is a rational number with a denominator of 1.

Because they are fractions, any rational number can also be expressed in decimal form. Any rational number can be represented as either:

1. a terminating decimal: $\frac{15}{8} = 1.875$, or
2. a repeating decimal: $\frac{4}{11} = 0.36363636\cdots = 0.\overline{36}$

We use a line drawn over the repeating block of numbers instead of writing the group multiple times.

Example 1.1.1: Writing Integers as Rational Numbers

Write each of the following as a rational number. Write a fraction with the integer in the numerator and 1 in the denominator.

a. 7
b. 0
c. -8

Solution

a. $7 = \frac{7}{1}$
b. $0 = \frac{0}{1}$
c. $-8 = \frac{-8}{1}$

Exercise 1.1.1

Write each of the following as a rational number.

a. 11
b. 3
c. -4

Answer available at the end of this section.

Example 1.1.2: Identifying Rational Numbers

Write each of the following rational numbers as either a terminating or repeating decimal.

a. $-\frac{5}{7}$
b. $\frac{15}{5}$
c. $\frac{13}{25}$

Solution

a. $-\frac{5}{7} = -0.714285714285\cdots = -0.\overline{714285}$, a repeating decimal
b. $\frac{15}{5} = 3$(or 3.0), a terminating decimal
c. $\frac{13}{25} = 0.52$, a terminating decimal

Exercise 1.1.2

Write each of the following rational numbers as either a terminating or repeating decimal.

a. $\frac{68}{17}$

b. $\frac{8}{13}$
c. $-\frac{13}{25}$

Answer available at the end of this section.

Irrational Numbers

At some point in the ancient past, someone discovered that not all numbers are rational numbers. A builder, for instance, may have found that the diagonal of a square with unit sides was not 2 or even 32,but was something else. Or a garment maker might have observed that the ratio of the circumference to the diameter of a roll of cloth was a little bit more than 3, but still not a rational number. Such numbers are said to be irrational because they cannot be written as fractions. These numbers make up the set of irrational numbers. Irrational numbers cannot be expressed as a fraction of two integers. It is impossible to describe this set of numbers by a single rule except to say that a number is irrational if it is not rational. So we write this as shown.

$$h \mid \text{h is not a rational number} \tag{1.1.2}$$

Example 1.1.3: Differentiating Rational and Irrational Numbers

Determine whether each of the following numbers is rational or irrational. If it is rational, determine whether it is a terminating or repeating decimal.

a. $\sqrt{25}$
b. $\frac{33}{9}$
c. $\sqrt{11}$
d. $\frac{17}{34}$
e. $0.3033033303333\ldots$

Solution

a. $\sqrt{25}$: This can be simplified as $\sqrt{25}=5$.Therefore,$\sqrt{25}$is rational.
b. $\frac{33}{9}$: Because it is a fraction,$\frac{33}{9}$is a rational number. Next, simplify and divide.

$$\frac{33}{9}=\frac{\cancel{33}}{\cancel{9}}$$

So,$\frac{33}{9}$ is rational and a repeating decimal.
c. $\sqrt{11}$: This cannot be simplified any further. Therefore,$\sqrt{11}$is an irrational number.
d. $\frac{17}{34}$: Because it is a fraction,$\frac{17}{34}$is a rational number. Simplify and divide.

$$\frac{17}{34}=0.5$$

So,$\frac{17}{34}$ is rational and a terminating decimal.
e. $0.3033033303333\ldots$ is not a terminating decimal. Also note that there is no repeating pattern because the group of $3s$ increases each time. Therefore it is neither a terminating nor a repeating decimal and, hence, not a rational number. It is an irrational number.

Exercise 1.1.3

Determine whether each of the following numbers is rational or irrational. If it is rational, determine whether it is a terminating or repeating decimal.

a. $\frac{7}{77}$
b. $\sqrt{81}$
c. $4.27027002700027\ldots$
d. $\frac{91}{13}$

e. $\sqrt{39}$

Answer available at the end of this section.

Real Numbers

Given any number n, we know that n is either rational or irrational. It cannot be both. The sets of rational and irrational numbers together make up the set of real numbers. As we saw with integers, the real numbers can be divided into three subsets: negative real numbers, zero, and positive real numbers. Each subset includes fractions, decimals, and irrational numbers according to their algebraic sign (+ or –). Zero is considered neither positive nor negative.

The real numbers can be visualized on a horizontal number line with an arbitrary point chosen as 0, with negative numbers to the left of 0 and positive numbers to the right of 0. A fixed unit distance is then used to mark off each integer (or other basic value) on either side of 0. Any real number corresponds to a unique position on the number line.The converse is also true: Each location on the number line corresponds to exactly one real number. This is known as a one-to-one correspondence. We refer to this as the real number line as shown in Figure$(1.1.1)$

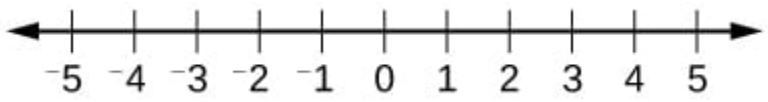


Figure $1.1.1$: The real number line.

Example $1.1.4$: Classifying Real Numbers

Classify each number as either positive or negative and as either rational or irrational. Does the number lie to the left or the right of 0 on the number line?

a. $-\frac{10}{3}$
b. $-\sqrt{5}$
c. -6π
d. $0.615384615384\ldots$

Solution

a. $-\frac{10}{3}$is negative and rational. It lies to the left of 0 on the number line.
b. $-\sqrt{5}$is positive and irrational. It lies to the right of 0.
c. $-\sqrt{289}=-\sqrt{17^2}=-17$ is negative and rational. It lies to the left of 0.
d. -6π is negative and irrational. It lies to the left of 0.
e. $0.615384615384\ldots$ is a repeating decimal so it is rational and positive. It lies to the right of 0.

Exercise $1.1.4$

Classify each number as either positive or negative and as either rational or irrational. Does the number lie to the left or the right of 0 on the number line?

a. $\sqrt{73}$
b. $-11.411411411\ldots$
c. $\frac{47}{19}$
d. $-\frac{\sqrt{5}}{2}$
e. 6.210735

Answer available at the end of this section.

Sets of Numbers as Subsets

Beginning with the natural numbers, we have expanded each set to form a larger set, meaning that there is a subset relationship between the sets of numbers we have encountered so far. These relationships become more obvious when seen as a diagram, such as Figure(1.1.2).

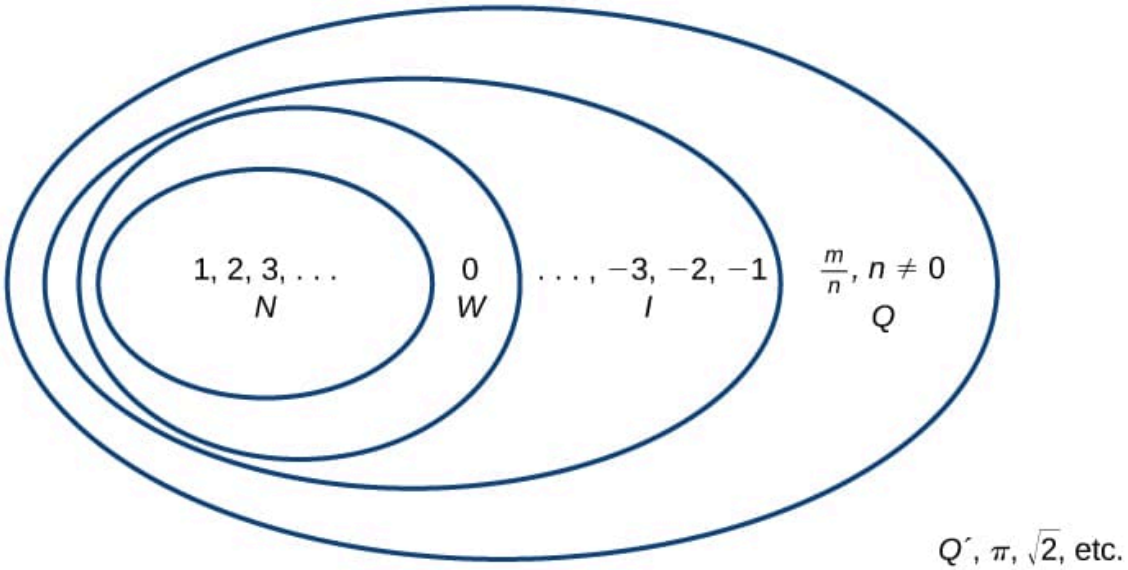


Figure 1.1.2: Sets of numbers. **N**: the set of natural numbers. **W**: the set of whole numbers. **I**: the set of integers. **Q**: the set of rational numbers. **Q´**: the set of irrational numbers

SETS OF NUMBERS

The set of natural numbers includes the numbers used for counting: $\{1, 2, 3, ...\}$.

The set of whole numbers is the set of natural numbers plus zero: $\{0, 1, 2, 3, ...\}$.

The set of integers adds the negative natural numbers to the set of whole numbers: $\{..., -3, -2, -1, 0, 1, 2, 3, ...\}$.

The set of rational numbers includes fractions written as $\{mn \parallel \text{m and n are integers and } neq 0\}$.

The set of irrational numbers is the set of numbers that are not rational, are nonrepeating, and are nonterminating: $\{h \parallel \text{h is not a rational number}\}$.

Example 1.1.5: Differentiating the Sets of Numbers

Classify each number as being a natural number (N), whole number (W), integer (I), rational number (Q), and/or irrational number (Q′).

a. $\sqrt{36}$
b. $\frac{8}{3}$
c. $\sqrt{73}$
d. -6
e. $3.2121121112\ldots$

Solution

	N	W	I	Q	Q'
a. $\sqrt{36} = 6$	X	X	X	X	
b. $\frac{8}{3} = 2.\overline{6}$				X	
c. $\sqrt{73}$					X
d. -6			X	X	
e. $3.2121121112...$					X

Exercise 1.1.5

Classify each number as being a natural number (N), whole number (W), integer (I), rational number (Q), and/or irrational number (Q').

a. $-\frac{35}{7}$
b. 0
c. $\sqrt{169}$
d. $\sqrt{24}$
e. $4.763763763...$

Answer available at the end of this section.

Interactive Exercise 1.1.5

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

View this Interactive Content online

Performing Calculations Using the Order of Operations

When we multiply a number by itself, we square it or raise it to a power of 2. For example, $4^2 = 4 \times 4 = 16$. We can raise any number to any power. In general, the exponential notation an means that the number or variable a is used as a factor n times.

$$a^n = a \cdot a \cdot a \cdots a \qquad \text{n factors}$$

In this notation, a^n is read as the nth power of a, where a is called the base and n is called the exponent. A term in exponential notation may be part of a mathematical expression, which is a combination of numbers and operations. For example, $24 + 6 \times \frac{2}{3} - 4^2$ is a mathematical expression.

To evaluate a mathematical expression, we perform the various operations. However, we do not perform them in any random order. We use the order of operations. This is a sequence of rules for evaluating such expressions.

Recall that in mathematics we use parentheses ( ), brackets [ ], and braces { } to group numbers and expressions so that anything

appearing within the symbols is treated as a unit. Additionally, fraction bars, radicals, and absolute value bars are treated as grouping symbols. When evaluating a mathematical expression, begin by simplifying expressions within grouping symbols.

The next step is to address any exponents or radicals. Afterward, perform multiplication and division from left to right and finally addition and subtraction from left to right.

Let's take a look at the expression provided.

$$24 + 6 \times \frac{2}{3} - 4^2$$

There are no grouping symbols, so we move on to exponents or radicals. The number 4 is raised to a power of 2, so simplify 4^2 as 16.

$$24 + 6 \times \frac{2}{3} - 4^2$$

$$24 + 6 \times \frac{2}{3} - 16$$

Next, perform multiplication or division, left to right.

$$24 + 6 \times \frac{2}{3} - 16$$

$$24 + 4 - 16$$

Lastly, perform addition or subtraction, left to right.

$$24 + 4 - 16$$

$$28 - 16$$

$$12$$

Therefore,

$$24 + 6 \times \frac{2}{3} - 4^2 = 12$$

For some complicated expressions, several passes through the order of operations will be needed. For instance, there may be a radical expression inside parentheses that must be simplified before the parentheses are evaluated. Following the order of operations ensures that anyone simplifying the same mathematical expression will get the same result.

ORDER OF OPERATIONS

Operations in mathematical expressions must be evaluated in a systematic order, which can be simplified using the acronym **PEMDAS**:

- **P**(arentheses)
- **E**(xponents)
- **M(**ultiplication) and **D**(ivision)
- **A**(ddition) and **S**(ubtraction)

HOW TO: Given a mathematical expression, simplify it using the order of operations.

1. Simplify any expressions within grouping symbols.
2. Simplify any expressions containing exponents or radicals.
3. Perform any multiplication and division in order, from left to right.
4. Perform any addition and subtraction in order, from left to right.

Example 1.1.6: Using the Order of Operations

Use the order of operations to evaluate each of the following expressions.

a. $(3 \times 2)^2 - 4 \times (6 + 2)$

b. $\dfrac{5^2 - 4}{7} - \sqrt{11 - 2}$

c. $6 - \mid 5 - 8 \mid + 3 \times (4 - 1)$

d. $\dfrac{14 - 3 \times 2}{2 \times 5 - 3^2}$

e. $7 \times (5 \times 3) - 2 \times [(6 - 3) - 4^2] + 1$

Solution

a.

$$\begin{aligned}
(3 \times 2)^2 - 4 \times (6 + 2) &= (6)^2 - 4 \times (8) && \text{Simplify parentheses} \\
&= 36 - 4 \times 8 && \text{Simplify exponent} \\
&= 36 - 32 && \text{Simplify multiplication} \\
&= 4 && \text{Simplify subtraction}
\end{aligned}$$

b.

$$\begin{aligned}
\frac{5^2 - 4}{7} - \sqrt{11 - 2} &= \frac{5^2 - 4}{7} - \sqrt{9} && \text{Simplify grouping symbols (radical)} \\
&= \frac{5^2 - 4}{7} - 3 && \text{Simplify radical} \\
&= \frac{25 - 4}{7} - 3 && \text{Simplify exponent} \\
&= \frac{21}{7} - 3 && \text{Simplify subtraction in numerator} \\
&= 3 - 3 && \text{Simplify division} \\
&= 0 && \text{Simplify subtraction}
\end{aligned}$$

Note that in the first step, the radical is treated as a grouping symbol, like parentheses. Also, in the third step, the fraction bar is considered a grouping symbol so the numerator is considered to be grouped.

c.

$$\begin{aligned}
6 - \mid 5 - 8 \mid + 3 \times (4 - 1) &= 6 - \mid -3 \mid + 3 \times 3 && \text{Simplify inside grouping symbols} \\
&= 6 - 3 + 3 \times 3 && \text{Simplify absolute value} \\
&= 6 - 3 + 9 && \text{Simplify multiplication} \\
&= 3 + 9 && \text{Simplify subtraction} \\
&= 12 && \text{Simplify addition}
\end{aligned}$$

d.

$$\begin{aligned}
\frac{14 - 3 \times 2}{2 \times 5 - 3^2} &= \frac{14 - 3 \times 2}{2 \times 5 - 9} && \text{Simplify exponent} \\
&= \frac{14 - 6}{10 - 9} && \text{Simplify products} \\
&= \frac{8}{1} && \text{Simplify differences} \\
&= 8 && \text{Simplify quotient}
\end{aligned}$$

In this example, the fraction bar separates the numerator and denominator, which we simplify separately until the last step.

e.

$$\begin{aligned}
7 \times (5 \times 3) - 2 \times [(6 - 3) - 4^2] + 1 &= 7 \times (15) - 2 \times [(3) - 4^2] + 1 && \text{Simplify inside parentheses} \\
&= 7 \times (15) - 2 \times (3 - 16) + 1 && \text{Simplify exponent} \\
&= 7 \times (15) - 2 \times (-13) + 1 && \text{Subtract} \\
&= 105 + 26 + 1 && \text{Multiply} \\
&= 132 && \text{Add}
\end{aligned}$$

Exercise 1.1.6

Use the order of operations to evaluate each of the following expressions.

a. $\sqrt{5^2 - 4^2} + 7 \times (5 - 4)^2$
b. $1 + \dfrac{7 \times 5 - 8 \times 4}{9 - 6}$
c. $|1.8 - 4.3| + 0.4 \times \sqrt{15 + 10}$
d. $\dfrac{1}{2} \times [5 \times 3^2 - 7^2] + \dfrac{1}{3} \times 9^2$
e. $[(3 - 8^2) - 4] - (3 - 8)$

Answer available at the end of this section.

Interactive Exercise 1.1.6

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

View this Interactive Content online

Example 1.1.7: Real-World Application

Fred earns $40 mowing lawns. He spends $10 on mp3s, puts a third of what is left in a savings account, and gets another $5 for washing his neighbor's car.

1. Write the numerical expression that represents the number of dollars Fred keeps (and does not put in his savings account). Remember the order of operations.
2. How much money does Fred keep?

Solution

1. $\frac{2}{3} \times (40 - 10) + 5$
2. $25

Interactive Exercise 1.1.7

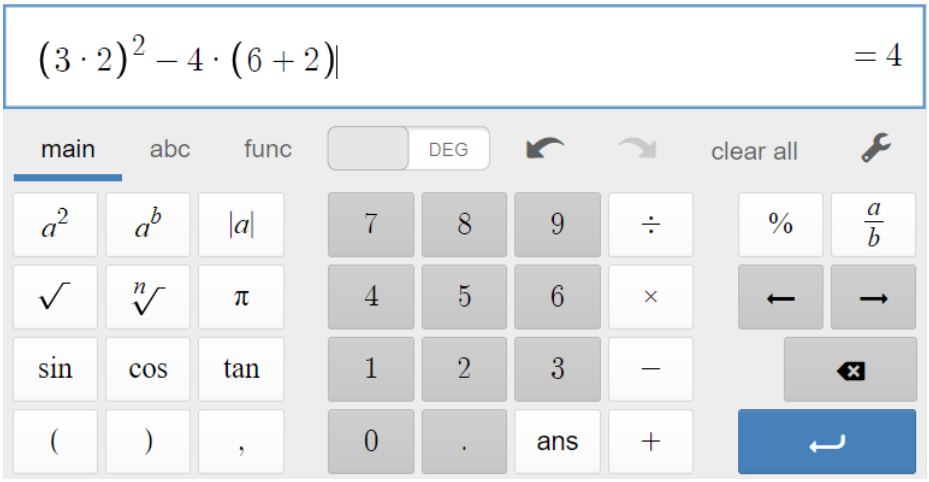
This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

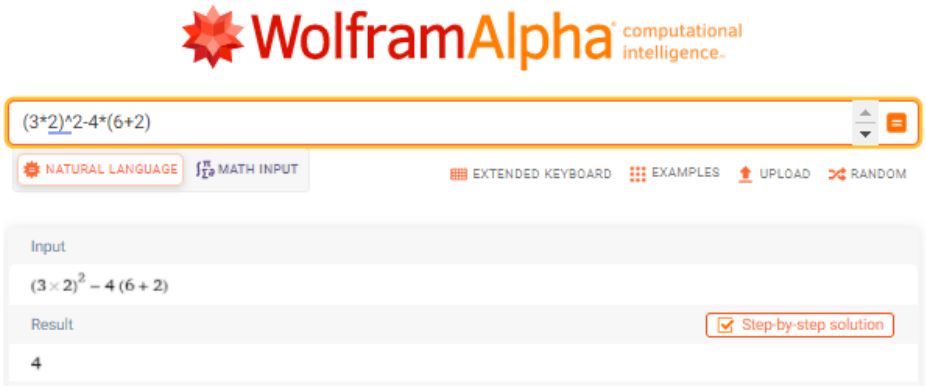
Computing Numerical Expressions with Technology

Luckily, when it comes to computing numerical expressions, we can trust technology with performing this task. Instead of following the order of operations and computing by hand, let's discuss how technology can be used.

In Desmos Scientific Calculator or Desmos Graphing Calculator, enter the expression then press Enter. For example, computing the numerical expression from part (a) of Example 1.1.6 is shown below:



In WolframAlpha, enter the expression then press Enter. For example, computing the numerical expression from part (a) of Example 1.1.6 is shown below:



Exercise 1.1.8: Using Technology to Evaluate Numerical Expressions

Compute each numerical expression.

a. $\dfrac{5+3}{5-3}$
b. $7 - 2(-2)$
c. $\dfrac{1}{3}\pi \cdot 11^2$

d. $((-2)^2 \cdot 3)^3$

e. $4(\frac{2}{3} - \frac{1}{3}) - 5(\frac{1}{3} - \frac{2}{3})$

Answer available at the end of this section.

Rounding Decimals

Decimals are another way of writing fractions whose denominators are powers of ten.

$$\begin{aligned} 0.1 &= \frac{1}{10} && \text{is “one tenth”} \\ 0.01 &= \frac{1}{100} && \text{is “one hundredth”} \\ 0.001 &= \frac{1}{1000} && \text{is “one thousandth”} \\ 0.0001 &= \frac{1}{10,000} && \text{is “one ten-thousandth”} \end{aligned} \tag{1.1.3}$$

Just as in whole numbers, each digit of a decimal corresponds to the place value based on the powers of ten. Figure shows the names of the place values to the left and right of the decimal point.

Place Value											
Hundred thousands	Ten thousands	Thousands	Hundreds	Tens	Ones		Tenths	Hundredths	Thousandths	Ten-thousandths	Hundred-thousandths
						.					

Figure 1.

When we work with decimals, it is often necessary to round the number to the nearest required place value. We summarize the steps for rounding a decimal here.

Rounding Decimals

1. Locate the given place value and mark it with an arrow.
2. Underline the digit to the right of the place value.
3. Is the underlined digit greater than or equal to 5?
   - Yes: add 1 to the digit in the given place value.
   - No: do not change the digit in the given place value
4. Rewrite the number, deleting all digits to the right of the rounding digit.

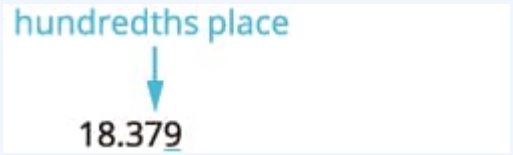
Example 1.1.9

Round 18.379 to the nearest ⓐ hundredth ⓑ tenth ⓒ whole number.

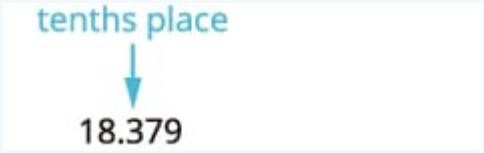
Solution

Round 18.379.

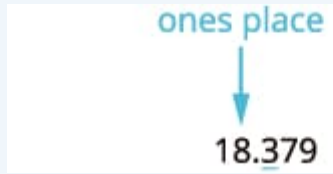
ⓐ to the nearest hundredth

Locate the hundredths place with an arrow.	
Underline the digit to the right of the given place value.	
Because 9 is greater than or equal to 5, add 1 to the 7.	
Rewrite the number, deleting all digits to the right of the rounding digit.	18.38
Notice that the deleted digits were NOT replaced with zeros.	So, 18.379 rounded to the nearest hundredth is 18.38.

ⓑ to the nearest tenth

Locate the tenths place with an arrow.	
Underline the digit to the right of the given place value.	
Because 7 is greater than or equal to 5, add 1 to the 3.	
Rewrite the number, deleting all digits to the right of the rounding digit.	18.4
Notice that the deleted digits were NOT replaced with zeros.	So, 18.379 rounded to the nearest tenth is 18.4.

ⓒ to the nearest whole number

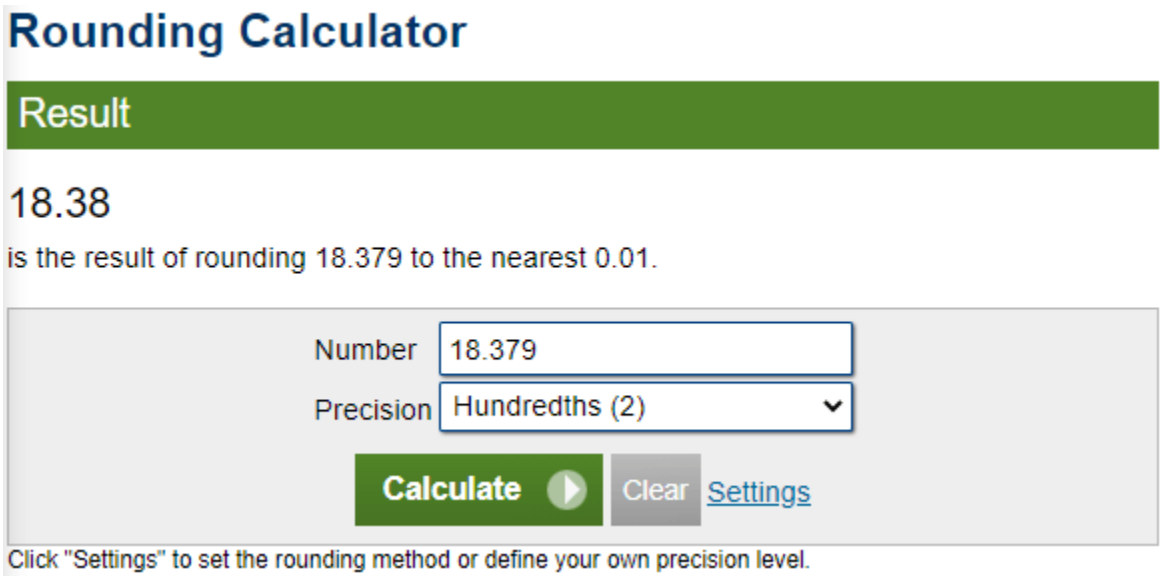
Locate the ones place with an arrow.	
Underline the digit to the right of the given place value.	
Since 3 is not greater than or equal to 5, do not add 1 to the 8.	
Rewrite the number, deleting all digits to the right of the rounding digit.	18
	So, 18.379 rounded to the nearest whole number is 18.

Exercise 1.1.9

Round 6.582 to the nearest ⓐ hundredth ⓑ tenth ⓒ whole number.

Answer available at the end of this section.

While rounding is a straightforward process feel free to check your work with any online calculator such as this one. For example, to round 18.379 to the nearest hundredth from part (1) of Example $1.1.9$, enter 18.379 and choose hundredths in the precision dropdown.



Interactive Exercise 1.1.9

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

View this Interactive Content online

Answers

Exercise 1.1.1

a. $\frac{11}{1}$
b. $\frac{3}{1}$
c. $-\frac{4}{1}$

Exercise 1.1.2

a. 4 (or 4.0), terminating
b. $0.\overline{615384}$, repeating
c. -0.85, terminating

Exercise 1.1.3

a. rational and terminating;

b. rational and repeating;

c. irrational

Exercise 1.1.4

a. positive, irrational
b. right negative, rational
c. left positive, rational
d. right negative, irrational
e. left positive, rational; right

Exercise 1.1.5

	N	W	I	Q	Q'
a. $-\frac{35}{7}$			X	X	
b. 0		X	X	X	
c. $\sqrt{169}$	X	X	X	X	
d. $\sqrt{24}$					X
e. $4.763763763\ldots$				X	

Exercise 1.1.6

a. 10
b. 2
c. 4.5
d. 25
e. -60

Exercise 1.1.8**: Using Technology to Evaluate Numerical Expressions**

a. 4
b. 11
c. $\frac{121}{3}\pi$
d. 1728
e. 3

Exercise 1.1.9

ⓐ 6.58 ⓑ 6.6 ⓒ 7

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

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

1.2: Percent

Learning Objectives

- Understand the definition and the meaning of percent notation.
- Convert between percent, fraction, and decimal notations.
- Use and understand percent notation in context

When one hears the word “percent,” other words come immediately to mind, words such as “century,” “cents,” or “centimeters.” A century equals 100 years. There are one hundred cents in a dollar and there are 100 centimeters in a meter. Thus, it should come as no surprise that percent means “parts per hundred.”

In the world we live in we are constantly bombarded with phrases that contain the word “percent.” The sales tax in California is 8.25%. An employee is asking his boss for a 5% raise. A union has seen a 6.25% increase in union dues. The population of a town is increasing at a rate of 2.25% per year.

In this section, we introduce the concept of percent, first addressing how to facilitate writing percents in fraction or decimal form and also performing the reverse operations, changing fractions and decimals to percents. Eventually, we apply this ability to solving common applications from the real world that use percents. We’ll tackle applications of commission and sales tax, discount and marked price, percent increase or decrease, and simple and compound interest.

Let’s begin the journey! In the square shown in Figure 1.2.1, a large square has been partitioned into ten rows of ten little squares in each row. In Figure 7.1, we’ve shaded 20 of 100 possible little squares, or 20% of the total number of little squares.

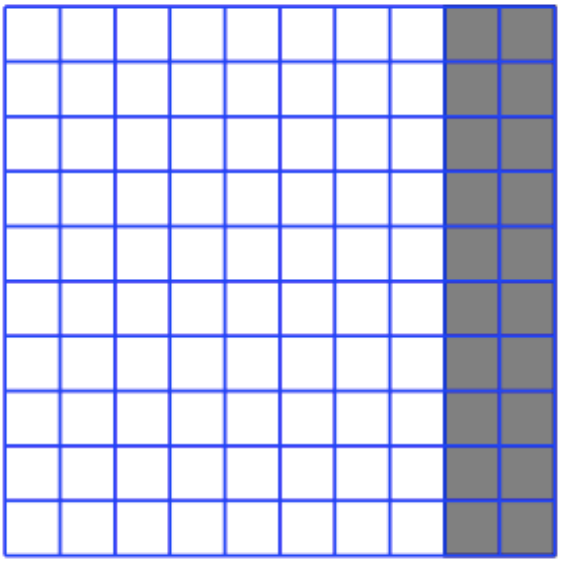


Figure 1.2.1: Shading 20 of 100 little squares, or 20% of the total number of little squares.

The Meaning of Percent

Percent means “parts per hundred.”

In Figure ({1{.}2{.}\eatSpaces{1} \endSpaces}\), 80 out of a possible 100 squares are left unshaded. Thus, 80% of the little squares are unshaded. If instead we shaded 35 out of the 100 squares, then 35% of the little squares would be shaded. If we shaded all of the little squares, then 100% of the little squares would be shaded (100 out of 100).

So, when you hear the word “percent,” think “parts per hundred.”

Changing a Percent to a Fraction

Based on the discussion above, it is fairly straightforward to change a percent to a fraction.

Percent to Fraction

To change a percent to a fraction, drop the percent sign and put the number over 100.

Example 1.2.1

Change 24% to a fraction.

Solution

Drop the percent symbol and put 24 over 100.

$$\begin{aligned} 24\% &= \frac{24}{100} && \text{Percent: Parts per hundred.} \\ &= \frac{6}{25} && \text{Reduce.} \end{aligned}$$

Hence, 24% = 6/25.

Exercise 1.2.1

Change 36% to a fraction reduced to lowest terms.

Answer available at the end of this section.

Interactive Exercise 1.2.1

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

View this Interactive Content online

Example 1.2.2

Change 28.4% to a fraction.

Solution

Drop the percent symbol and put 28.4 over 100.

$$\begin{aligned} 28.4\% &= \frac{28.4}{100} && \text{Percent: Parts per hundred.} \\ &= \frac{28.4 \cdot 10}{100 \cdot 10} && \text{Multiply numerator and denominator by 10.} \\ &= \frac{284}{1000} && \text{Multiplying by 10 moves decimal point one place right.} \\ &= \frac{71 \cdot 4}{250 \cdot 4} && \text{Factor.} \\ &= \frac{71}{250} && \text{Cancel common factor.} \end{aligned}$$

Exercise 1.2.2

Change 87.5% to a fraction reduced to lowest terms.

Answer available at the end of this section.

Exercise 1.2.2

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

View this Interactive Content online



Changing a Percent to a Decimal

To change a percent to a decimal, we need only remember that percent means “parts per hundred.”

Example 1.2.3

Change 23.25% to a decimal.

Solution

Drop the percent symbol and put 23.25 over 100.

$$\begin{aligned} 23.25\% &= \frac{23.25}{100} && \text{Percent: Parts per hundred.} \\ &= 0.2325 && \text{Dividing by 100 moves decimal point 2 places left.} \end{aligned}$$

Therefore, 23.25% = 0.2325.

Exercise 1.2.3

Change 2.4% to a decimal.

Answer available at the end of this section.

This last example motivates the following simple rule.

Changing a Percent to a Decimal

To change a percent to a decimal, drop the percent symbol and move the decimal point two places to the left.

Interactive Exercise 1.2.3

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

View this Interactive Content online

Changing a Decimal to a Percent

Changing a decimal to a percent is the exact opposite of changing a percent to a decimal. In the latter case, we drop the percent symbol and move the decimal point 2 places to the left. The following rule does just the opposite.

Changing a Decimal to a Percent

To change a decimal to a percent, move the decimal point two places to the right and add a percent symbol.

Example 1.2.4

Change 0.0725 to a percent.

Solution

Move the decimal point two places to the right and add a percent symbol.

$$\begin{aligned} 0.0725 &= 007.25\% \\ &= 7.25\% \end{aligned}$$

Exercise 1.2.4

Change to 0.0375 to a percent.

Answer available at the end of this section.

Interactive Exercise 1.2.4

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

View this Interactive Content online

Example 1.2.5

Change 1.025 to a percent.

Solution

Move the decimal point two places to the right and add a percent symbol.

$$\begin{aligned} 1.025 &= 102.5\% \\ &= 102.5\% \end{aligned}$$

Exercise 1.2.5

Change 2.525 to a percent.

Answer available at the end of this section.

Interactive Exercise 1.2.5

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

View this Interactive Content online

Changing a Fraction to a Percent

One way to proceed is to first change the fraction to a decimal, then change the resulting decimal to a percent.

Fractions to Percents: Technique #1

To change a fraction to a percent, follow these steps:

1. Divide the numerator by the denominator to change the fraction to a decimal.
2. Move the decimal point in the result two places to the right and append a percent symbol.

Example 1.2.6

Use Technique #1 to change 5/8 to a percent.

Solution

Change 5/8 to a decimal, then change the decimal to a percent.

To change 5/8 to a decimal, divide 5 by 8. Since the denominator is a product of twos, the decimal should terminate.

$$\begin{array}{r} 0.625 \\ 8\overline{)5.000} \\ \underline{4\,8} \\ 20 \\ \underline{16} \\ 40 \\ \underline{40} \\ 0 \end{array}$$

To change 0.625 to a percent, move the decimal point 2 places to the right and append a percent symbol.

$$0.625 = 0\ 62.5\% = 62.5\%$$

Exercise 1.2.6

Change 5/16 to a percent.

Answer available at the end of this section.

Interactive Exercise 1.2.6

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

View this Interactive Content online

A second technique is to create an equivalent fraction with a denominator of 100.

Fractions to Percents: Technique #2

To change a fraction to a percent, create an equivalent fraction with a denominator of 100. This technique should only be used when the denominator of the fraction is a divisor of 100.

Example 1.2.7

Use Technique #2 to change 3/5 to a percent.

Solution

Note that the denominator 5 is a divisor of 100, that is 100=5*20. To create an equivalent fraction for 3/5 with a denominator of 100 multiple both, the numerator and denominator, by 20:

$$\frac{3}{5} = \frac{3 \cdot 20}{5 \cdot 20} = \frac{60}{100} \tag{1.2.1}$$

Thus,

$$\frac{3}{5} = 60\%.$$

Exercise 1.2.7

Change 4/25 to a percent.

Answer available at the end of this section.

Interactive Exercise 1.2.7

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

View this Interactive Content online

Sometimes we will be content with an approximation.

Example 1.2.8

Change 4/13 to a percent. Round your answer to the nearest tenth of a percent.

Solution

We will use Technique #1.

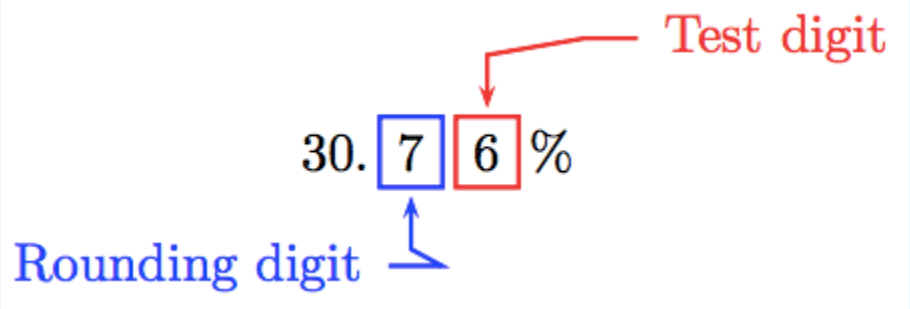
To change 4/13 to a decimal, divide 4 by 13. Since the denominator has factors other than 2's and 5's, the decimal will repeat. However, we intend to round to the nearest tenth of a percent, so we will carry the division to four decimal places only. (*Four places are necessary because we will be moving the decimal point two places to the right.*)

$$
\begin{array}{r}
0.625 \\
\hline
8\overline{)5.000} \\
4\,8 \\
\hline
20 \\
16 \\
\hline
40 \\
40 \\
\hline
0
\end{array}
$$

To change the decimal to a percent, move the decimal point two places to the right.

$$0.3076 \approx 0\ 30.76\% \approx 30.76\%$$

To round to the nearest tenth of a percent, identify the rounding and test digits.



Because the test digit is greater than or equal to 5, add 1 to the rounding digit and truncate. Thus,

$$0.03076 \approx 30.8\%.$$

Exercise 1.2.8

Change 4/17 to a percent. Round your answer to the nearest tenth of a percent.

Answer available at the end of this section.

Interactive Exercise 1.2.8

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

View this Interactive Content online

Using Percent in Context

Example 1.2.9: Major Hurricanes

5 of the 8 hurricanes in 2008 were categorized as major. Write the fractional number of major hurricanes in 2008 as a percent. *NOAA Associated Press 5/22/09*

Solution

$$\frac{5}{8} = 0.625 = 62.5\%$$

Exercise 1.2.9: Major Hurricanes

The 2021 hurricane season produced seven hurricanes—Elsa, Grace, Henri, Ida, Larry, Nicholas and Sam. Four hurricanes—Grace, Ida, Larry and Sam—reached major hurricane status (Category 3 and above). Ida and Sam reached Category 4 status. (Source: Facts + Statistics: Hurricanes | III) Write the fractional number of major hurricanes in 2021 as a percent.

Answer available at the end of this section.

Interactive Exercise 1.2.9

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

View this Interactive Content online

Answers

Exercise $1.2.1$

9/25

Exercise $1.2.2$

7/8

Exercise $1.2.3$

0.024

Exercise $1.2.4$

3.75%

Exercise $1.2.5$

252.5%

Exercise $1.2.6$

31.35%

Exercise $1.2.7$

16%

Exercise $1.2.8$

23.5%

Exercise $1.2.9$**: Major Hurricanes**

$\frac{4}{7} = 0.5714 = 57.14\%$

This page titled 1.2: Percent is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by David Arnold.

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

1.3: Algebra Essentials

Learning Objectives

- Evaluate and simplify algebraic expressions.
- Solve an equation
- Use a formula

So far, the mathematical expressions we have seen have involved real numbers only. In mathematics, we may see expressions such as $x+5$, $\frac{4}{3}\pi r^3$, or $\sqrt{2m^3n^2}$. In the expression $x+5$, 5 is called a **constant** because it does not vary and x is called a **variable** because it does. (In naming the variable, ignore any exponents or radicals containing the variable.) An **algebraic expression** is a collection of constants and variables joined together by the algebraic operations of addition, subtraction, multiplication, and division.

We have already seen some real number examples of exponential notation, a shorthand method of writing products of the same factor. When variables are used, the constants and variables are treated the same way.

$$(-3)^5 = (-3)\times(-3)\times(-3)\times(-3)\times(-3) \Rightarrow x^5 = x\times x\times x\times x\times x$$
$$(2\times 7)^3 = (2\times 7)\times(2\times 7)\times(2\times 7) \quad \Rightarrow (yz)^3 = (yz)\times(yz)\times(yz)$$

In each case, the exponent tells us how many factors of the base to use, whether the base consists of constants or variables.

Evaluating Algebraic Expressions

Any variable in an algebraic expression may take on or be assigned different values. When that happens, the value of the algebraic expression changes. To evaluate an algebraic expression means to determine the value of the expression for a given value of each variable in the expression. Replace each variable in the expression with the given value, then simplify the resulting expression using the order of operations. If the algebraic expression contains more than one variable, replace each variable with its assigned value and simplify the expression as before.

Example 1.3.1: Describing Algebraic Expressions

List the constants and variables for each algebraic expression.

a. $x+5$
b. $\frac{4}{3}\pi r^3$
c. $\sqrt{2m^3n^2}$

Solution

	Constants	Variables
a. $x+5$	5	x
b. $\frac{4}{3}\pi r^3$	$\frac{4}{3}$, π	r
c. $\sqrt{2m^3n^2}$	2	m,n

Exercise 1.3.1: Describing Algebraic Expressions

List the constants and variables for each algebraic expression.

a. $2\pi r(r+h)$
b. $2(L+W)$
c. $4y^3+y$

Answer available at the end of this section.

Example 1.3.2: Evaluating an Algebraic Expression at Different Values

Evaluate the expression $2x-7$ for each value for x.

a. $x=0$
b. $x=1$
c. $x=12$
d. $x=-4$

Solution

a. Substitute 0 for x.

$$\begin{aligned} 2x-7 &= 2(0)-7 \\ &= 0-7 \\ &= -7 \end{aligned}$$

b. Substitute 1 for x.

$$\begin{aligned} 2x-7 &= 2(1)-7 \\ &= 2-7 \\ &= -5 \end{aligned}$$

c. Substitute $\frac{1}{2}$ for x.

$$\begin{aligned} 2x-7 &= 2\left(\frac{1}{2}\right)-7 \\ &= 1-7 \\ &= -6 \end{aligned}$$

d. Substitute -4 for x.

$$\begin{aligned} 2x-7 &= 2(-4)-7 \\ &= -8-7 \\ &= -15 \end{aligned}$$

Exercise 1.3.2: Evaluating an Algebraic Expression at Different Values

Evaluate the expression $11-3y$ for each value for y.

a. $y=2$
b. $y=0$
c. $y=\frac{2}{3}$
d. $y=-5$

Answer available at the end of this section.

Interactive Exercise 1.3.2

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

View this Interactive Content online

Example 1.3.3: Evaluating Algebraic Expressions

Evaluate each expression for the given values.

a. $x + 5$ for $x = -5$

b. $\dfrac{t}{2t - 1}$ for $t = 10$

c. $\dfrac{4}{3}\pi r^3$ for $r = 5$

d. $a + ab + b$ for $a = 11$, $b = -8$

e. $\sqrt{2m^3 + n^2}$ for $m = 2$, $n = 3$

Solution

a. Substitute -5 for x.

$$\begin{aligned} x+5 &= (-5)+5 \\ &= 0 \end{aligned}$$

b. Substitute 10 for t.

$$\begin{aligned} \frac{t}{2t-1} &= \frac{(10)}{2(10)-1} \\ &= \frac{10}{20-1} \\ &= \frac{10}{19} \end{aligned}$$

c. Substitute 5 for r.

$$\begin{aligned} \frac{4}{3}\pi r^3 &= \frac{4}{3}\pi(5)^3 \\ &= \frac{4}{3}\pi(125) \\ &= \frac{500}{3}\pi \end{aligned}$$

d. Substitute 11 for a and -8 for b.

$$\begin{aligned} a+ab+b &= (11)+(11)(-8)+(-8) \\ &= 11-88-8 \\ &= -85 \end{aligned}$$

e. Substitute 2 for m and 3 for n.

$$\begin{aligned} \sqrt{2m^3+n^2} &= \sqrt{2(2)^3+(3)^2} \\ &= \sqrt{2(8)+(9)} \\ &= \sqrt{25} \\ &= 5 \end{aligned}$$

Interactive Exercise 1.3.3

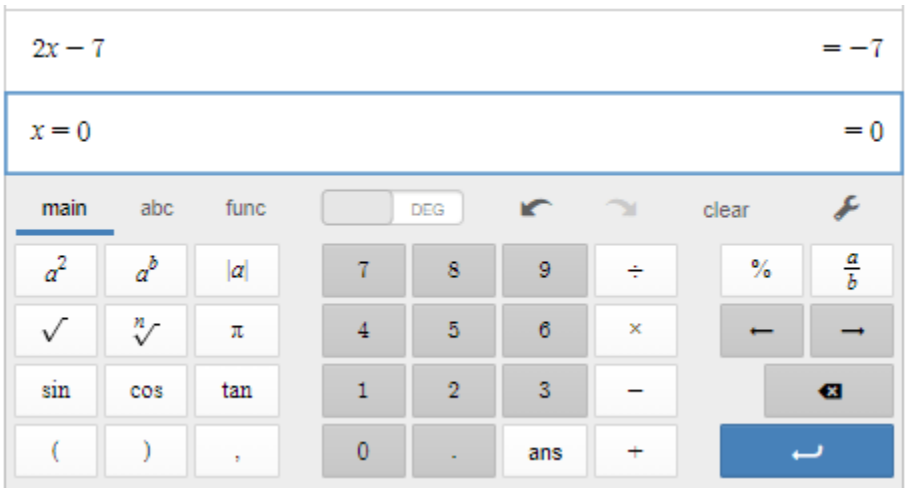
This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Evaluating Algebraic Expressions with Technology

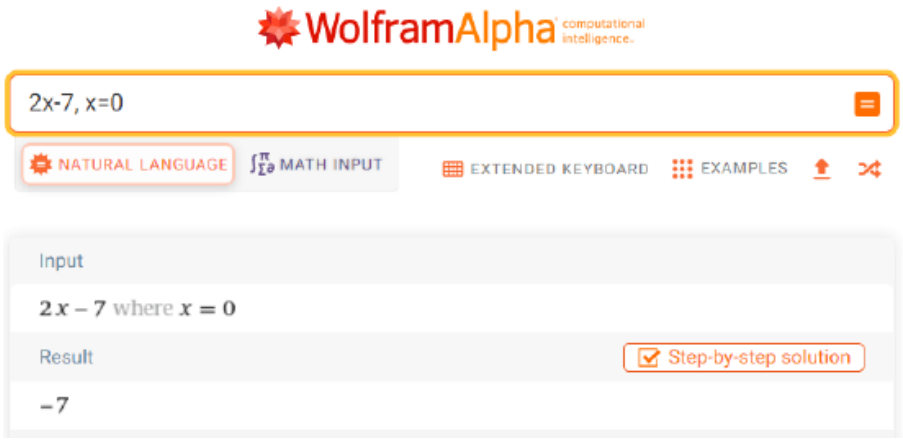
Luckily, when it comes to evaluating expressions, we can trust technology with performing this task. Instead of plugging in the values and computing by hand, let's discuss how technology can be used.

In Desmos Scientific Calculator, enter the expression then press Enter and enter $x =$ followed by the value at which the expression is being evaluated. In case there are multiple variables involved, repeat the previous step for every additional variable. Once the values for all involved variables are entered the ⚠ notification will be replaced with the value of the expression. For example, evaluating $2x - 7$ at $x = 0$ from part (a) of Example $1.3.2$ is shown below:



In WolframAlpha, enter the expression followed by the value at which the expression is being evaluated. In case there are multiple variables involved, repeat the previous step for every additional variable. Once the values for all involved variables are entered

press Enter and the value of the expression will appear among the results. For example, evaluating $2x - 7$ at $x = 0$ from part (a) of Example $1.3.2$ is shown below:



Exercise 1.3.4: Using Technology to Evaluate Algebraic Expressions

Evaluate each expression for the given values.

a. $\dfrac{y+3}{y-3}$ for $y = 5$
b. $7 - 2t$ for $t = -2$
c. $\dfrac{1}{3}\pi r^2$ for $r = 11$
d. $(p^2q)^3$ for $p = -2$, $q = 3$
e. $4(m-n) - 5(n-m)$ for $m = \dfrac{2}{3}$ $n = \dfrac{1}{3}$

Answer available at the end of this section.

Solving Equations

An **equation** is an expression that is *equal* to another expression. Unlike expressions, equations have an equal sign.

Example 1.3.5

Examples of equations:

a) $x - 4 = 6$

b) $5x - 6 = 4x + 2$

c) $6x - 30 = 0$

d) $x^2 + 3x - 4 = 0$

e) $3x^2 - 2x = -1$

f) $x^3 + x^2 + x + 1 = 0$

g) $2x - 5 > 3$ is not an equation but rather an inequality

A **solution of an equation** is any value of the variable that satisfies the equality, that is, it makes the Left-Hand Side (LHS) and the Right-Hand Side (RHS) of the equation the same value.

Example 1.3.6

Solutions of equations:

a) A solution for $x - 4 = 6$ is $x = 10$ because the LHS evaluated at $x = 10$ is $10 - 4 = 6$ which is equal to the RHS.

b) A solution for $5x - 6 = 4x + 2$ is $x = 8$ because the LHS evaluated at $x = 8$ is $5(8) - 6 = 40 - 6 = 34$ and the RHS evaluated at $x = 8$ is $4x + 2 = 4(8) + 2 = 32 + 2 = 34$, and they are equal!

So, given a value of x, we can check if it is a solution or not by evaluating simultaneously the LHS and RHS of an equation. If they are **equal**, then the value is **a solution**. If they are **not equal**, then the value is **not a solution**.

Example 1.3.7: Checking Solutions

a) Is $x = 2$ a solution of the equation

$$-4x + 8 + x = 5 - 2x + 1$$

?

The LHS evaluated at $x = 2$ is $-4(2) + 8 + 2 = -8 + 8 + 2 = 2$.

The RHS evaluated at $x = 2$ is $5 - 2(2) + 1 = 5 - 4 + 1 = 5 + (-4) + 1 = 1 + 1 = 2$.

Since they are equal, then we say that $x = 2$ is a solution for the given equation.

b) Is $x = -1$ a solution of the equation

$$x^2 + 4x = -3x + 8$$

?

The LHS evaluated at $x = -1$ is $(-1)^2 + 4(-1) = 1 + (-4) = -3$.

The RHS evaluated at $x = -1$ is $-3(-1) + 8 = 3 + 8 = 11$.

Since $-3 \neq 11$, then we say that $x = -1$ is **not a solution** for the given equation.

c) Is $x = -2$ a solution to

$$x^2 - 2x + 1 = 3x^2 + 2x + 1$$

?

The LHS evaluated at $x = -2$ is $(-2)^2 - 2(-2) + 1 = 4 + 4 + 1 = 9$.

The RHS evaluated at $x = -2$ is $3(-2)^2 + 2(-2) + 1 = 3 \cdot 4 - 4 + 1 = 12 - 4 + 1 = 9$. since $9 = 9$, the $LHS = RHS$ and $x = -2$ is a solution to the equation.

Exercise 1.3.7: Checking Solutions

Check whether $x = -6$ is a solution of the equation

$$10 + 10x = 13 + 6(x + 1)?$$

Answer available at the end of this section.

Interactive Exercise 1.3.7

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

View this Interactive Content online

To **solve an equation** means to **find all the solution(s)** for that equation. The method to solve an equation depends on the kind of equation at hand. For example, we solve linear equations by isolating the variable and quadratic equations by using the quadratic formula.

Example 1.3.8

Solve the equation $3x - 3 = 2x + 4$ for x.

Solution

We will isolate all terms containing an x on the left side of this equation (we could just as well have chosen to isolate terms containing x on the right side of the equation). To this end, we don't want the –3 on the left side of the equation (we want it on the right), so we add 3 to both sides of the equation and simplify.

$$\begin{aligned} 3x - 3 &= 2x + 4 \\ 3x - 3 + 3 &= 2x + 4 + 3 \\ 3x &= 2x + 7 \end{aligned} \tag{1.3.1}$$

Remember that we have chosen to isolate all terms containing x on the left side of the equation. So, for our next step, we choose to subtract 2x from both sides of the equation (this will "move" it from the right over to the left), then simplify.

$$\begin{aligned} 3x &= 2x + 7 \\ 3x - 2x &= 2x + 7 - 2x \\ x &= 7 \end{aligned} \tag{1.3.2}$$

To check the solution, substitute x = 7 in the original equation to obtain

$$\begin{aligned} 3x - 3 &= 2x + 4 \\ 3(7) - 3 &= 2(7) + 4 \\ 21 - 3 &= 14 + 4 \\ 18 &= 18 \end{aligned} \tag{1.3.3}$$

The last line is a true statement, so $x = 7$ checks and is a solution of $3x - 3 = 2x + 4$.

Example 1.3.9

Solve the equation $x + 3 = x^2 - 2$ for x.

Solution

Collect all the terms on the left side:

$$x^2 - x - 5 = 0 \tag{1.3.4}$$

Apply the quadratic formula:

$$\begin{aligned} x &= \frac{-(-1) \pm \sqrt{(-1)^2 - 4 \cdot 1 \cdot (-5)}}{2} \\ &= \frac{1 \pm \sqrt{21}}{2} \end{aligned} \tag{1.3.5}$$

Interactive Exercise 1.3.8

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

View this Interactive Content online

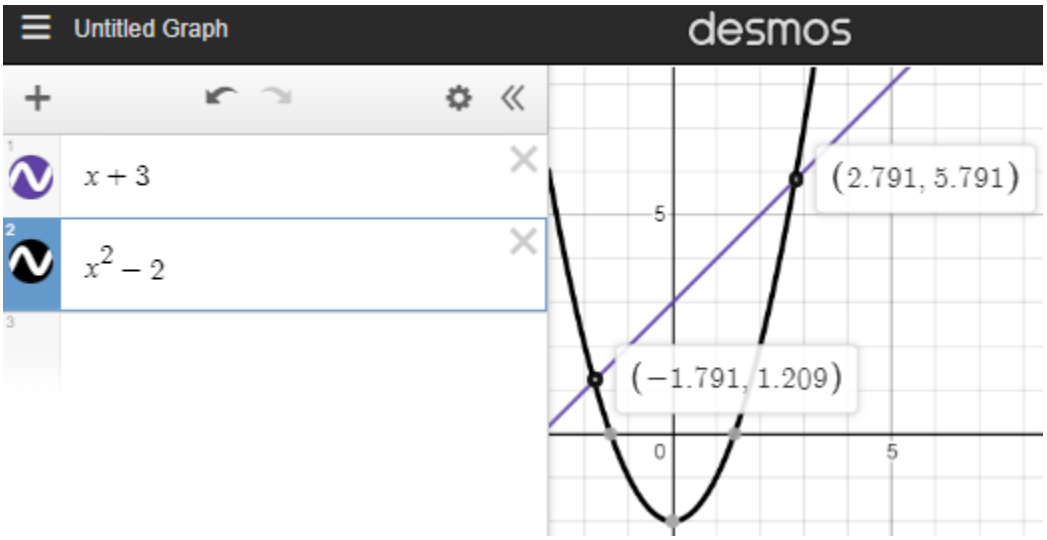
Solving Equations with Technology

Luckily, when it comes to solving equations, we can trust technology with performing this task. Instead of isolating the variable by hand or using the quadratic formula, let's discuss how technology can be used.

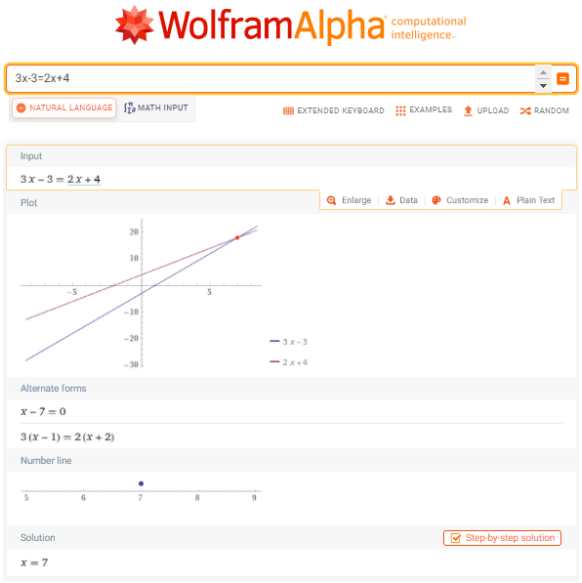
In Desmos Graphing Calculator, enter the expression on the left side of the equation then press Enter and enter the expression on the right side of the equation. The graphs will appear on the coordinate plane representing the sides of the equation. The x-coordinates of the points of intersection are the solutions of the equation. For example, solving the equation $3x - 3 = 2x + 4$ in Example $1.3.8$ is shown below:



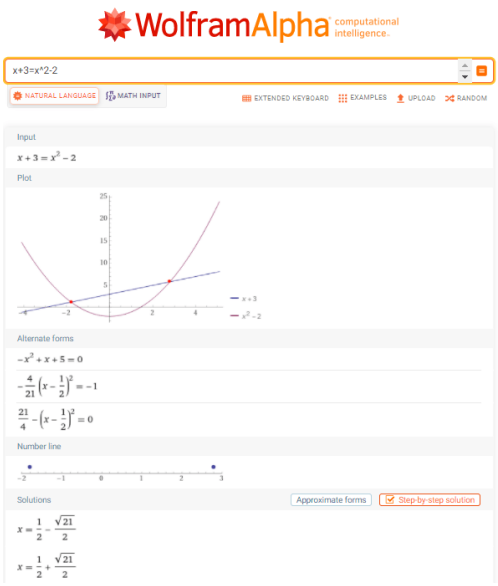
For example, solving the equation $x + 3 = x^2 - 2$ in Example $1.3.9$ is shown below:



In WolframAlpha, enter the equation and press Enter. Along with many other things associated with the equation you will find the solution at the very bottom of the output page. For example, solving the equation $3x - 3 = 2x + 4$ in Example $1.3.8$ is shown below:



For example, solving the equation $x+3=x^2-2$ in Example $1.3.9$ is shown below:



The main difference between WolframAlpha and Desmos is that WolframAlpha can provide the exact solutions (in case of solving quadratic equation $x+3=x^2-2$, $x=1/2\pm\frac{\sqrt{21}}{2}$) while Desmos can only do an approximation (in case of solving quadratic equation $x+3=x^2-2$, $x=-1.791$ or 2.791). WolframAlpha in general is a more powerful tool but for graphing purposes, Desmos is great enough.

Using Formulas

A **formula** is an equation expressing a relationship between constant and variable quantities. Very often, the equation is a means of finding the value of one quantity (often a single variable) in terms of another or other quantities. In other words, a formula can be converted to an equation in one variable if the values are provided for all but one variable. One of the most common examples is the formula for finding the area A of a circle in terms of the radius r of the circle: $A=\pi r^2$. For any value of r, the area A can be found by evaluating the expression πr^2. On the other hand, for any value of A, the radius r can be found by solving the equation. For example, if $A=6$ then the quadratic equation that needs to be solved is $\pi r^2=6$.

Example 1.3.10: Using a Formula

A right circular cylinder with radius r and height h has the surface area S (in square units) given by the formula $S = 2\pi r(r + h)$. See Figure 1.3.7. Find the surface area of a cylinder with radius 6 in. and height 9 in. Leave the answer in terms of π.

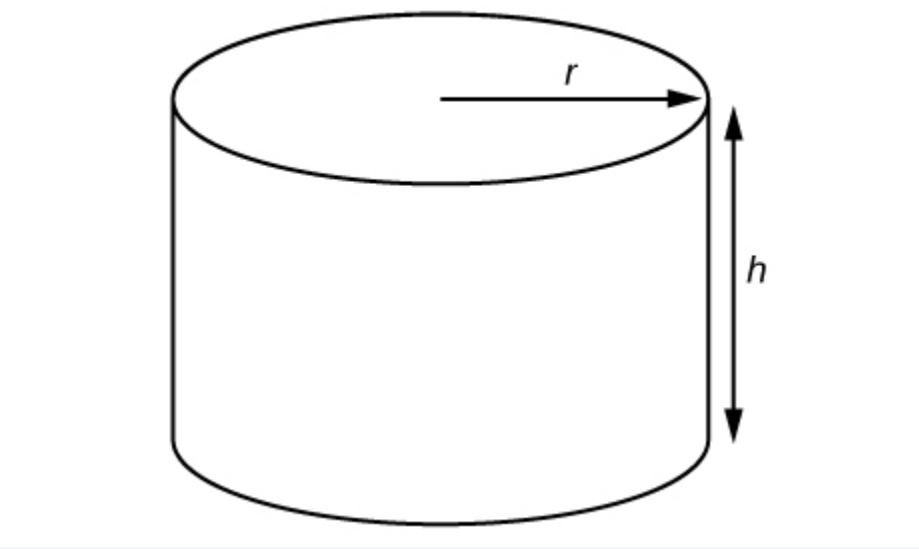


Figure 1.3.7

Solution

Evaluate the expression $2\pi r(r + h)$ for $r = 6$ and $h = 9$.

$$\begin{aligned} S &= 2\pi r(r + h) \\ &= 2\pi(6)[(6) + (9)] \\ &= 2\pi(6)(15) \\ &= 180\pi \end{aligned}$$

The surface area is 180π square inches.

Exercise 1.3.10: Using a Formula

A right circular cylinder with radius r and height h has the surface area S (in square units) given by the formula $S = 2\pi r(r + h)$. See Figure 1.3.7. Find the surface area of a cylinder with radius 2 in. and height 5 in. Leave the answer in terms of π.

Answer available at the end of this section.

Interactive Exercise 1.3.10

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

View this Interactive Content online

Here are some more examples of applications that come from geometry.

Example 1.3.11: Using Formulas

a) Find the perimeter of a rectangle with length 33cm and width 17cm.

$$\begin{aligned} P &= 2l + 2w \\ &= 2(33\text{cm}) + 2(17\text{cm}) \\ &= 66\text{cm} + 34\text{cm} \\ &= 100\text{cm} \end{aligned}$$

b) Find the area of a rectangle that is 12cm long and 16cm wide.

$$\begin{aligned} A &= l \cdot w \\ &= 12\text{cm} \cdot 16\text{cm} \\ &= 192\text{cm}^2 \end{aligned}$$

c) Find the area of a triangle with height of 20in and a base of 30in.

$$\begin{aligned} A &= \frac{1}{2} b \cdot h \\ &= \frac{1}{2} 20\text{in} \cdot 30\text{in} \\ &= 300\text{in}^2 \end{aligned}$$

d) Find the area of a circle with radius 7cm, round your answer to the nearest tenth.

$$\begin{aligned} A &= \pi r^2 \\ &= \pi (7\text{cm})^2 \\ &\approx 3.14 \cdot 49\text{cm}^2 \\ &= 153.86\text{cm}^2 \\ &\approx 153.9\text{cm}^2 \end{aligned}$$

Example 1.3.12

A photograph with length L and width W is placed in a matte of width 8 centimeters (cm). The area of the matte (in square centimeters, or cm^2 is found to be $A = (L + 16)(W + 16) - L \cdot W$. See Figure $1.3.8$.

1. Find the area of a matte for a photograph with length 32 cm and width 24 cm.
2. Find the length of a photograph with the area of a matte 1376 sq cm and width 30 cm.
3. Find the width of a photograph with the area of a matte 928 sq cm and length 18 cm.

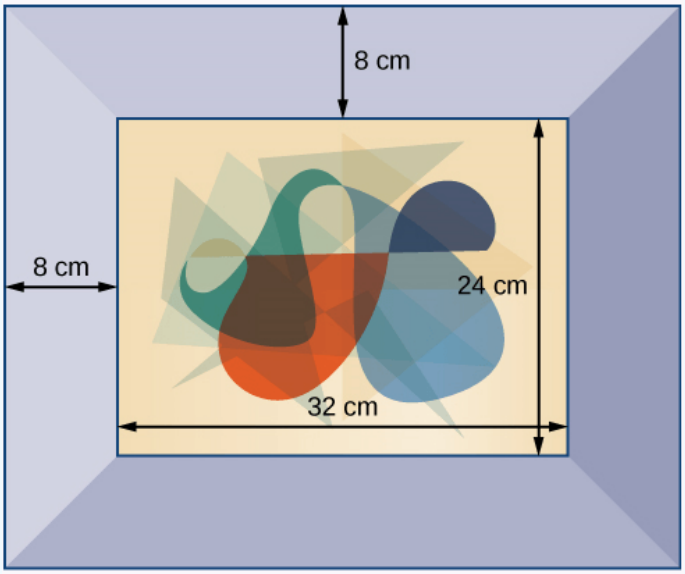


Figure $1.3.8$

Solution

1. Evaluate the expression $A = (L + 16)(W + 16) - L \cdot W$ for $L = 32$ and $W = 24$:
$$\begin{aligned} A &= (32 + 16)(24 + 16) - 32 \cdot 24 \\ &= 48 \cdot 40 - 32 \cdot 24 \\ &= 1152\text{cm}^2 \end{aligned}$$
2. Solve the equation $A = (L + 16)(W + 16) - L \cdot W$ for $A = 1376$ and $W = 30$:

$$1376 = (L+16)(30+16) - L \cdot 30$$
$$1376 = 16L + 736$$
$$L = 40\text{cm}$$

3. Solve the equation $A = (L+16)(W+16) - L \cdot W$ for $A = 928$ and $L = 18$:

$$928 = (18+16)(W+16) - 18 \cdot W$$
$$928 = 16W + 544$$
$$W = 24\text{cm}$$

Exercise 1.3.12: Using a Formula

A photograph with length L and width W is placed in a matte of width 10 centimeters (cm). The area of the matte (in square centimeters, or cm^2 is found to be $A = (L+20)(W+20) - L \cdot W$. See Figure 1.3.8.

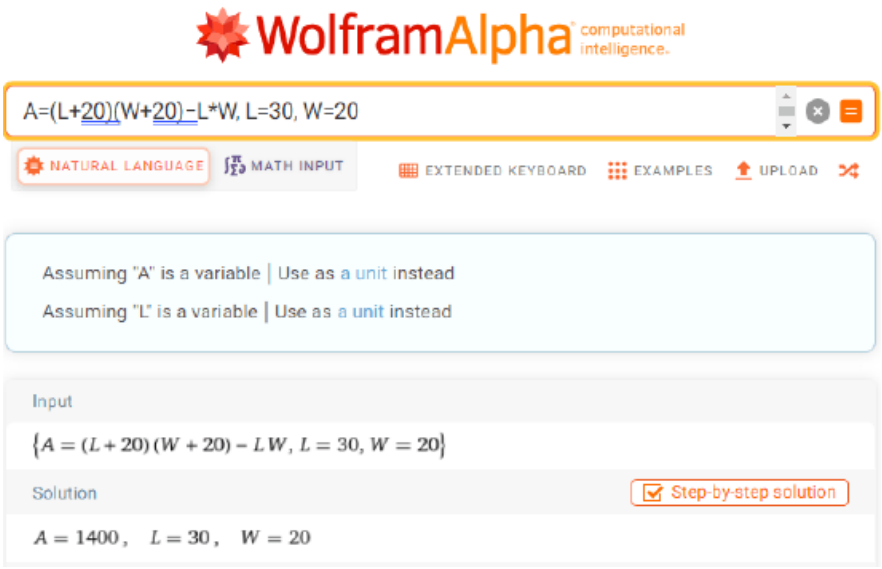
1. Find the area of a matte for a photograph with length 30 cm and width 20 cm.
2. Find the length of a photograph with the area of a matte 1600 sq cm and width 10 cm.
3. Find the width of a photograph with the area of a matte 928 sq cm and length 18 cm.

Answer available at the end of this section.

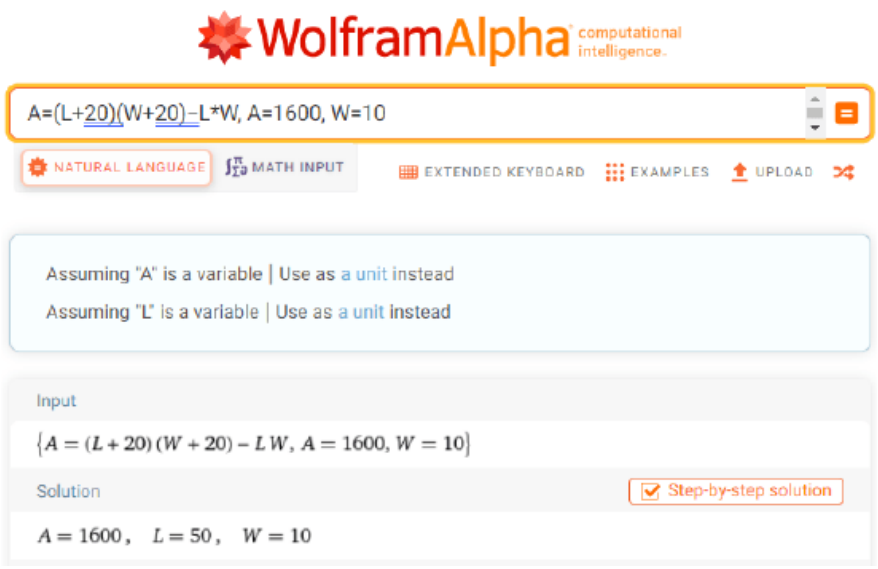
Using Formulas with Technology

Luckily, when it comes to using formulas, we can trust technology with performing this task. Instead of plugging in the values by hand, and then evaluating an expression or solving an equation let's discuss how technology can be used.

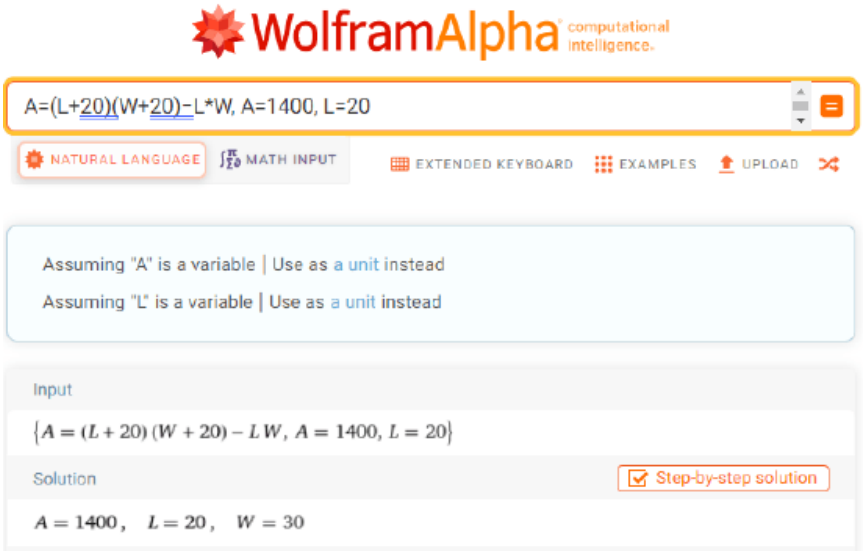
In WolframAlpha, enter the formula, followed by the variables and their assigned values separated by comma. Once the values for all involved variables are entered press Enter and WolframAlpha will determine the value of the missing variable. For example, using the formula $A = (L+20)(W+20) - L \cdot W$ for $L = 30$ and $W = 20$ in part 1 of Example 1.3.11 is shown below:



Similarly, using the formula $A = (L+20)(W+20) - L \cdot W$ for $A = 1600$ and $W = 10$ in part 2 of Example 1.3.11 is shown below:



Finally, using the formula $A = (L + 20)(W + 20) - L \cdot W$ for $A = 1400$ and $L = 20$ in part 3 of Example $1.3.11$ is shown below:



Interactive Exercise $1.3.12$

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

View this Interactive Content online

Answers

Exercise 1.3.1**: Describing Algebraic Expressions**

	Constants	Variables
a. $2\pi r(r+h)$	$2,\pi$	r,h
b. $2(L+W)$	2	L, W
c. $4y^3+y$	4	y

Exercise 1.3.2**: Evaluating an Algebraic Expression at Different Values**

a. 5

b. 11

c. 9

d. 26

Exercise 1.3.4**: Using Technology to Evaluate Algebraic Expressions**

a. 4
b. 11
c. $\frac{121}{3}\pi$
d. 1728
e. 3

Exercise 1.3.7**: Checking Solutions**

No, since the LHS is $10+10(-6)=-50$ but the RHS is $13+6(-6+1)=-17$.

Exercise 1.3.10**: Using a Formula**

Evaluate the expression $2\pi r(r+h)$ for $r=2$ and $h=5$.

$$\begin{aligned} S &= 2\pi r(r+h) \\ &= 2\pi(2)[(2)+(5)] \\ &= 2\pi(2)(7) \\ &= 28\pi \end{aligned}$$

The surface area is 28π square inches.

Exercise $1.3.12$**: Using a Formula**

1. Evaluate the expression $A = (L + 20)(W + 20) - L \cdot W$ for $L = 30$ and $W = 20$:
$$\begin{aligned} A &= (30 + 20)(20 + 20) - 30 \cdot 20 \\ &= 50 \cdot 40 - 30 \cdot 20 \\ &= 1400\text{cm}^2 \end{aligned}$$
2. Solve the equation $A = (L + 20)(W + 20) - L \cdot W$ for $A = 1600$ and $W = 10$:
$$\begin{aligned} 1600 &= (L + 20)(10 + 20) - L \cdot 10 \\ 1600 &= 20L + 600 \\ L &= 50\text{cm} \end{aligned}$$
3. Solve the equation $A = (L + 20)(W + 20) - L \cdot W$ for $A = 1400$ and $L = 20$:
$$\begin{aligned} 1400 &= (20 + 20)(W + 20) - 20 \cdot W \\ 1400 &= 20W + 800 \\ W &= 30\text{cm} \end{aligned}$$

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

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

1.4: Graphing

Learning Objectives

- Setup and use Cartesian coordinate system.
- Graph equations in two variables.

René Descartes (1596-1650) was a French philosopher and mathematician. As a philosopher, he is famous for the saying "Cogito ergo sum" ("I think, therefore I am"), and his writings led many to consider him the *Father of Modern Philosophy*. Even today, a number of his writings are standard faire in university philosophy departments.

However, it is Descartes' work in mathematics that forms the basis for this chapter, particularly his invention of the *Cartesian Coordinate System* which bears his name. Descartes' invention of the coordinate system created an entirely new branch of mathematics called *analytic geometry*, which established a permanent link between the plane and solid geometry of the ancient Greeks and the algebra and analysis of modern mathematics. As a result of his work, mathematicians were able to describe curves with equations, unheard of before Descartes' invention of the coordinate system. Rather than describing a circle as the "locus of all points equidistant from a given point," mathematicians were now able to refer to a circle centered at the point (0, 0) with radius r as the graph of the equation $x^2 + y^2 = r^2$.

The bridge created between geometry and analysis as a result of Descartes' methods laid the groundwork for the discovery of the calculus by Newton and Leibniz. For his efforts, mathematicians often refer to Descartes as the *Father of Analytic Geometry*.

In this section, we will introduce readers to the Cartesian coordinate system and explain the correspondence between points in the plane and ordered pairs of numbers. Once an understanding of the coordinate system is sufficiently developed, we will develop the concept of the graph of an equation.

We begin with the definition of an *ordered pair*.

Ordered Pair

The construct (x, y), where x and y are any real numbers, is called an ordered pair of real numbers.

$(4, 3)$, $(-3, 4)$, $(-2, -3)$, and $(3, -1)$ are examples of ordered pairs.

Order Matters

Pay particular attention to the phrase "ordered pairs." Order matters. Consequently, the ordered pair (x, y) is not the same as the ordered pair (y, x), because the numbers are presented in a different order.

The Cartesian Coordinate System

Pictured in Figure $1.4.1$ is a Cartesian Coordinate System. On a grid, we've created two real lines, one horizontal labeled x (we'll refer to this one as the x-axis), and the other vertical labeled y (we'll refer to this one as the y-axis).

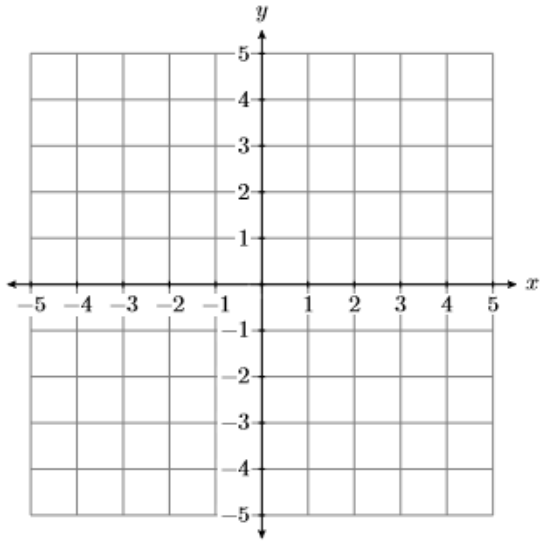


Figure 1.4.1: The Cartesian Coordinate System.

Two Important Points:

Here are two important points to be made about the horizontal and vertical axes in Figure 1.4.1.

1. As you move from left to right along the horizontal axis (the x-axis in Figure 1.4.1), the numbers grow larger. The positive direction is to the right, the negative direction is to the left.
2. As you move from bottom to top along the vertical axis (the y-axis in Figure 1.4.1), the numbers grow larger. The positive direction is upward, the negative direction is downward.

Additional Comments:

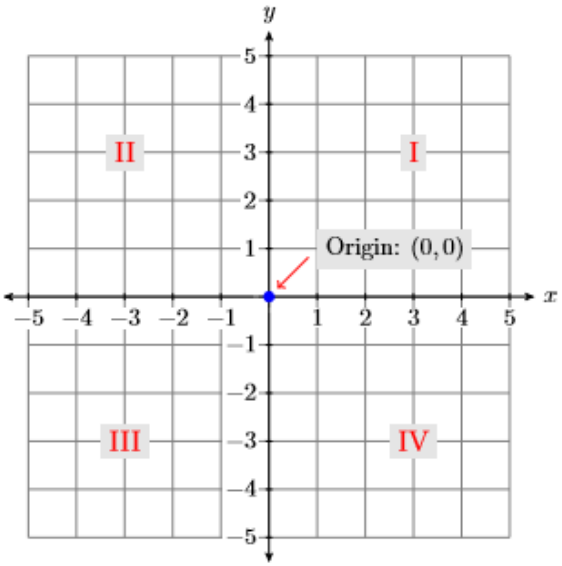


Figure 1.4.2: Numbering the quadrants and indicating the coordinates of the origin.

Two additional comments are in order:

3. The point where the horizontal and vertical axes intersect in Figure 1.4.2 is called the origin of the coordinate system. The origin has coordinates $(0, 0)$.
4. The horizontal and vertical axes divide the plane into four quadrants, numbered $\mathrm{I}, \mathrm{II}, \mathrm{MI},$ and IV (roman numerals for one,

two, three, and four), as shown in Figure 1.4.2. Note that the quadrants are numbered in a counter-clockwise order.

Note

Rene Descartes (1596-1650) was a French philosopher and mathematician who is well known for the famous phrase“cogito ergo sum” (I think, therefore I am), which appears in his Discours de la methode pour bien conduire sa raison, et chercher la verite dans les sciences (Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences). In that same treatise, Descartes introduces his coordinate system, a method for representing points in the plane via pairs of real numbers. Indeed, the Cartesian plane of modern day is so named in honor of Rene Descartes, who some call the “Father of Modern Mathematics”

Plotting Ordered Pairs

Before we can plot any points or draw any graphs, we first need to set up a Cartesian Coordinate System on a sheet of graph paper? How do we do this? What is required?

How to setup Cartesian coordinate system

Draw and label each axis.

If we are going to plot points (x,y), then, on a sheet of graph paper, perform each of the following initial tasks.

1. Use a ruler to draw the horizontal and vertical axes.
2. Label the horizontal axis as the x-axis and the vertical axis as the y-axis.

We don’t always label the horizontal axis as the x-axis and the vertical axis as the y-axis. For example, if we want to plot the velocity of an object as a function of time, then we would be plotting points (t, v). In that case, we would label the horizontal axis as the t-axis and the vertical axis as the v-axis.

Indicate the scale on each axis.

3. Label at least one vertical gridline with its numerical value.
4. Label at least one horizontal gridline with its numerical value.

The scales on the horizontal and vertical axes may differ. However, on each axis, the scale must remain consistent. That is, as you count to the right from the origin on the x-axis, if each gridline represents one unit, then as you count to the left from the origin on the x-axis, each gridline must also represent one unit. Similar comments are in order for the y-axis, where the scale must also be consistent, whether you are counting up or down.

The result of this first step is shown in Figure 1.4.3.

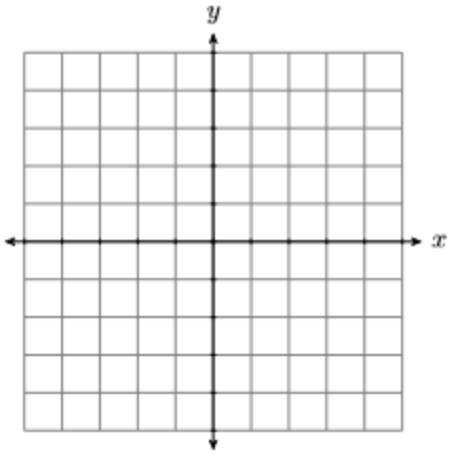


Figure 1.4.3: Draw and label each axis.

An example is shown in Figure 1.4.4. Note that the scale indicated on the x-axis indicates that each gridline counts as 1-unit as we count from left-to-right. The scale on the y-axis indicates that each gridlines counts as 2-units as we count from bottom-to-top.

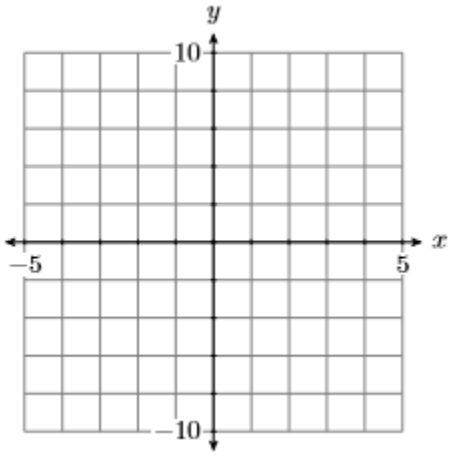


Figure 1.4.4: Indicate the scale on each axis.

Now that we know how to set up a Cartesian Coordinate System on a sheet of graph paper, here are two examples of how we plot points on our coordinate system.

To plot the ordered pair $(4, 3)$, start at the origin and move 4 units to the right along the horizontal axis, then 3 units upward in the direction of the vertical axis.

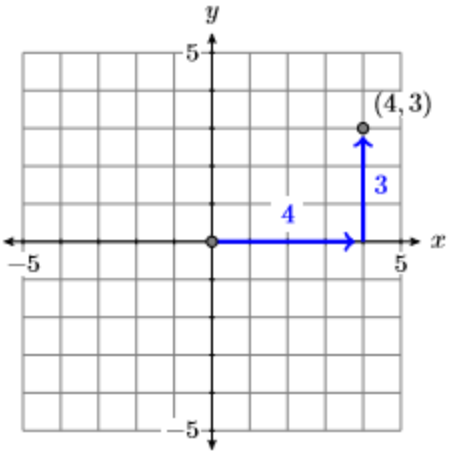


Figure 1.4.5: Plot of $(4, 3)$.

To plot the ordered pair $(-2, -3)$, start at the origin and move 2 units to the left along the horizontal axis, then 3 units downward in the direction of the vertical axis.

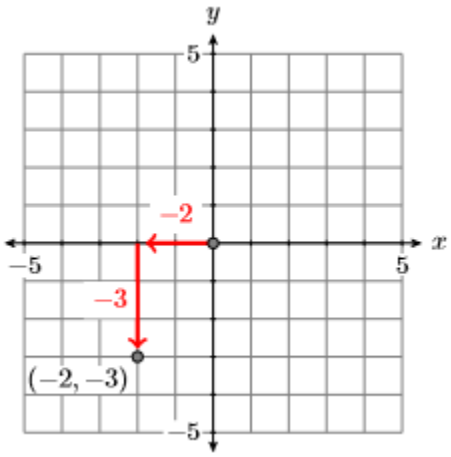


Figure 1.4.6: Plot of $(-2, -3)$

Interactive Exercise

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

View this Interactive Content online

Continuing in this manner, each ordered pair (x, y) of real numbers is associated with a unique point in the Cartesian plane. Vice-versa, each point in the Cartesian point is associated with a unique ordered pair of real numbers. Because of this association, we begin to use the words "point" and "ordered pair" as equivalent expressions, sometimes referring to the "point" (x, y) and other times to the "ordered pair" (x, y).

Example 1.4.1

Identify the coordinates of the point P in Figure 1.4.7.

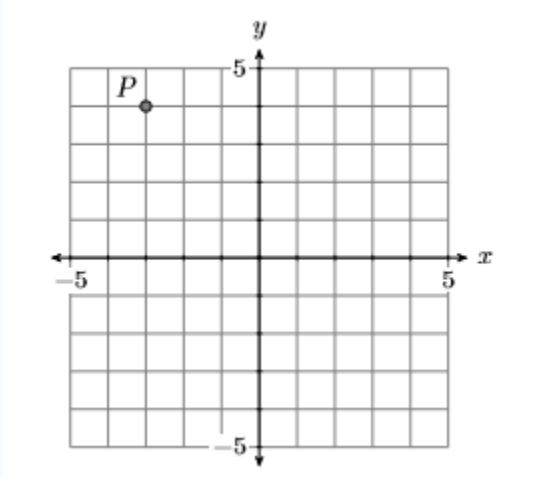


Figure 1.4.7: Identify the coordinates of the point P.

Solution

In Figure 1.4.8, start at the origin, move 3 units to the left and 4 units up to reach the point P. This indicates that the coordinates of the point P are $(-3, 4)$.

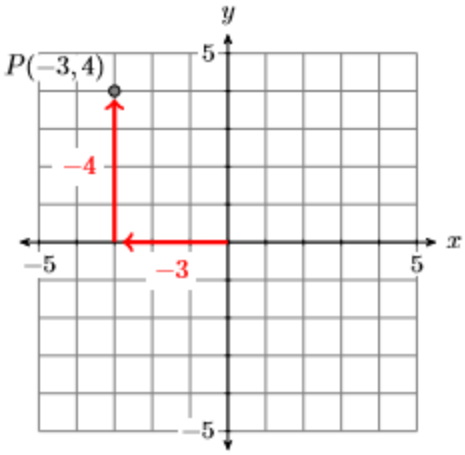
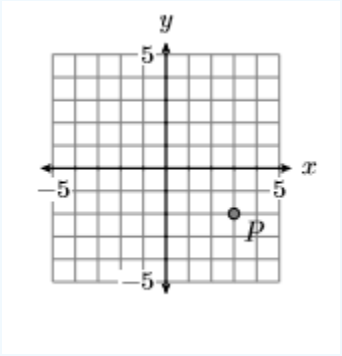


Figure 1.4.8: Start at the origin, move 3 units left and 4 units up.

Exercise 1.4.1

Identify the coordinates of the point P in the graph below.



Answer available at the end of this section.

Interactive Exercise 1.4.1

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

View this Interactive Content online

Equations in Two Variables

Note

The variables do not have to always be x and y. For example, the equation $v = 2 + 3.2t$ is an equation in two variables, v and t.

The equation $y = x + 1$ is an equation in two variables, in this case x and y. Consider the point $(x, y) = (2, 3)$. If we substitute 2 for x and 3 for y in the equation $y = x + 1$, we get the following result:

$$\begin{aligned} y &= x + 1 \quad \text{Original equation.} \\ 3 &= 2 + 1 \quad \text{Substitute: } 2 \text{ for } x, 3 \text{ for } y \\ 3 &= 3 \quad \text{Simplify both sides.} \end{aligned}$$

Because the last line is a true statement, we say that $(2, 3)$ is a solution of the equation $y = x + 1$. Alternately, we say that $(2, 3)$ satisfies the equation $y = x + 1$. On the other hand, consider the point $(x, y) = (-3, 1)$. If we substitute -3 for x and 1 for y in the equation $y = x + 1$, we get the following result.

$$\begin{aligned} y &= x + 1 \quad \text{Original equation.} \\ 1 &= -3 + 1 \quad \text{Substitute: } -3 \text{ for } x, 1 \text{ for } y \\ 1 &= -2 \quad \text{Simplify both sides.} \end{aligned}$$

Because the last line is a false statement, the point $(-3, 1)$ is **not** a solution of the equation $y = x + 1$; that is, the point $(-3, 1)$ does **not** satisfy the equation $y = x + 1$.

Solutions of an equation in two variables

Given an equation in the variables x and y and a point $(x, y) = (a, b)$, if upon subsituting a for x and b for y a true statement results, then the point $(x, y) = (a, b)$ is said to be a solution of the given equation. Alternately, we say that the point $(x, y) = (a, b)$ satisfies the given equation.

Example 1.4.2

Which of the ordered pairs $(0, -3)$ and $(1, 1)$ satisfy the equation $y = 3x - 2$?

Solution

Substituting the ordered pairs $(0, -3)$ and $(1, 1)$ into the equation $y = 3x - 2$ lead to the following results:

Consider $(x, y) = (0, -3)$. Substitute 0 for x and -3 for y:

$$\begin{aligned} y &= 3x - 2 \\ -3 &= 3(0) - 2 \\ -3 &= -2 \end{aligned}$$

The resulting statement is false.

Consider $(x, y) = (1, 1)$. Substitute 1 for x and 1 for y:

$$\begin{aligned} y &= 3x - 2 \\ 1 &= 3(1) - 2 \\ 1 &= 1 \end{aligned}$$

The resulting statement is true.

Thus, the ordered pair $(0, -3)$ does not satisfy the equation $y = 3x - 2$, but the ordered pair $(1, 1)$ does satisfy the equation $y = 3x - 2$.

Exercise 1.4.2

Which of the ordered pairs $(-1, 3)$ and $(2, 1)$ satisfy the equation $y = 2x + 5$?

Answer available at the end of this section.

Interactive Exercise 1.4.2

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

View this Interactive Content online

Graphing Equations in Two Variables

Let's first define what is meant by the graph of an equation in two variables.

The graph of an equation

The graph of an equation is the set of all points that satisfy the given equation.

Example 1.4.3

Sketch the graph of the equation $y = x + 1$.

Solution

The definition requires that we plot all points in the Cartesian Coordinate System that satisfy the equation $y = x + 1$. Let's first create a table of points that satisfy the equation. Start by creating three columns with headers x, y, and (x, y), then select some values for x and put them in the first column.

Take the first value of x, namely $x = -3$, and substitute it into the equation $y = x + 1$.

$$\begin{aligned} y &= x + 1 \\ y &= -3 + 1 \\ y &= -2 \end{aligned}$$



Thus, when $x = -3$, we have $y = -2$. Enter this value into the table.

Continue substituting each tabular value of x into the equation $y = x + 1$ and use each result to complete the corresponding entries in the table.

$$\begin{aligned} y &= -3 + 1 = -2 \\ y &= -2 + 1 = -1 \\ y &= -1 + 1 = 0 \\ y &= 0 + 1 = 1 \\ y &= 1 + 1 = 2 \\ y &= 2 + 1 = 3 \\ y &= 3 + 1 = 4 \end{aligned}$$



The last column of the table now contains seven points that satisfy the equation $y = x + 1$. Plot these points on a Cartesian Coordinate System (see Figure \(\PageIndex{9}\)).

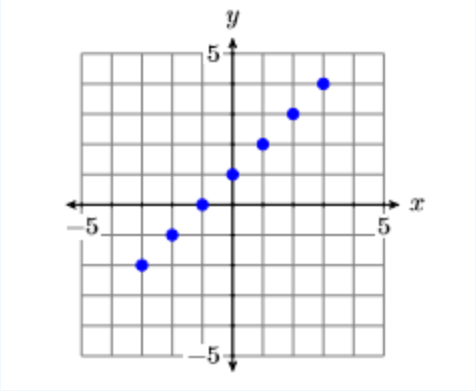


Figure 1.4.9: Seven points that satisfy the equation $y = x + 1$.

In Figure 1.4.9, we have plotted seven points that satisfy the given equation $y = x + 1$. However, the definition requires that we plot all points that satisfy the equation. It appears that a pattern is developing in Figure 1.4.9, but let's calculate and plot a few more points in order to be sure. Add the x-values -2.5, -1.5, -0.5, 0.5, 1.5, and 2.5 to the x-column of the table, then use the equation $y = x + 1$ to evaluate y at each one of these x-values.

$$y = -2.5 + 1 = -1.5$$
$$y = -1.5 + 1 = -0.5$$
$$y = -0.5 + 1 = 0.5$$
$$y = 0.5 + 1 = 1.5$$
$$y = 1.5 + 1 = 2.5$$
$$y = 2.5 + 1 = 3.5$$



Add these additional points to the graph in Figure 1.4.9 to produce the image shown in Figure 1.4.10.

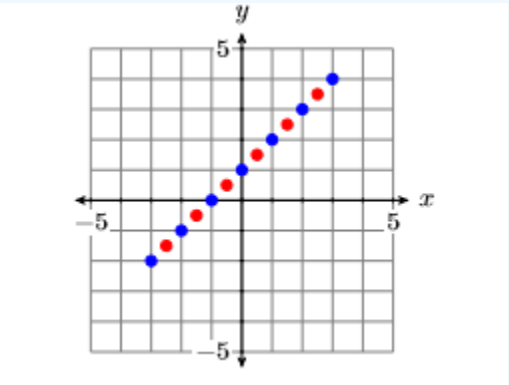


Figure 1.4.10: Adding additional points to the graph of $y = x + 1$.

There are an infinite number of points that satisfy the equation $y = x + 1$. In Figure 1.4.10, we've plotted only 13 points that satisfy the equation. However, the collection of points plotted in Figure 1.4.10 suggest that if we were to plot the remainder of the points that satisfy the equation $y = x + 1$, we would get the graph of the line shown in Figure 1.4.11.

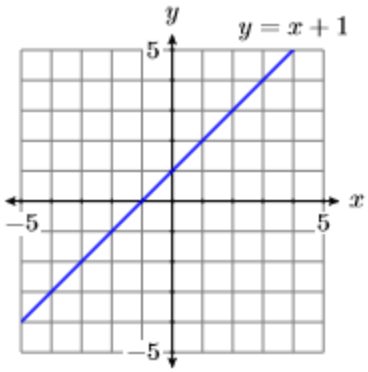


Figure 1.4.11: The graph of $y = x + 1$ is a line.

Exercise 1.4.3

Sketch the graph of the equation $y = -x + 2$.

Answer available at the end of this section.

Interactive Exercise 1.4.3

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

View this Interactive Content online

Guidelines and Requirements

Example 1.4.3 suggests that we should use the following guidelines when sketching the graph of an equation.

Guidelines for drawing the graph of an equation

When asked to draw the graph of an equation, perform each of the following steps:

1. Set up and calculate a table of points that satisfy the given equation.
2. Set up a Cartesian Coordinate System on graph paper and plot the points in your table on the system. Label each axis (usually x and y) and indicate the scale on each axis.
3. If the number of points plotted are enough to envision what the shape of the final curve will be, then draw the remaining points that satisfy the equation as imagined. Use a ruler if you believe the graph is a line. If the graph appears to be a curve, freehand the graph without the use of a ruler.
4. If the number of plotted points do not provide enough evidence to envision the final shape of the graph, add more points to your table, plot them, and try again to envision the final shape of the graph. If you still cannot predict the eventual shape of the graph, keep adding points to your table and plotting them until you are convinced of the final shape of the graph.

Here are some additional recommendations for sketching the graph of an equation.

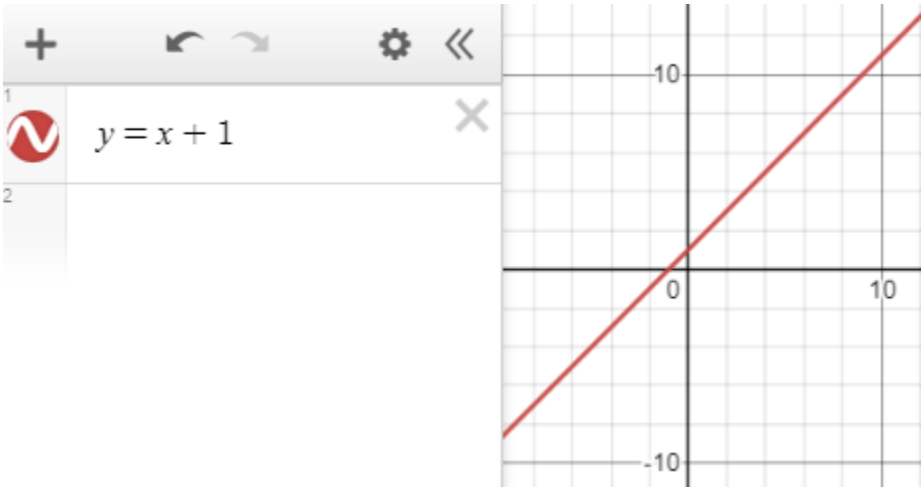
Graph paper, lines, curves, and rulers.

5. All graphs are to be drawn on graph paper.
6. All lines are to be drawn with a ruler. This includes the horizontal and vertical axes.
7. If the graph of an equation is a curve instead of a line, then the graph should be drawn freehand, without the aid of a ruler.

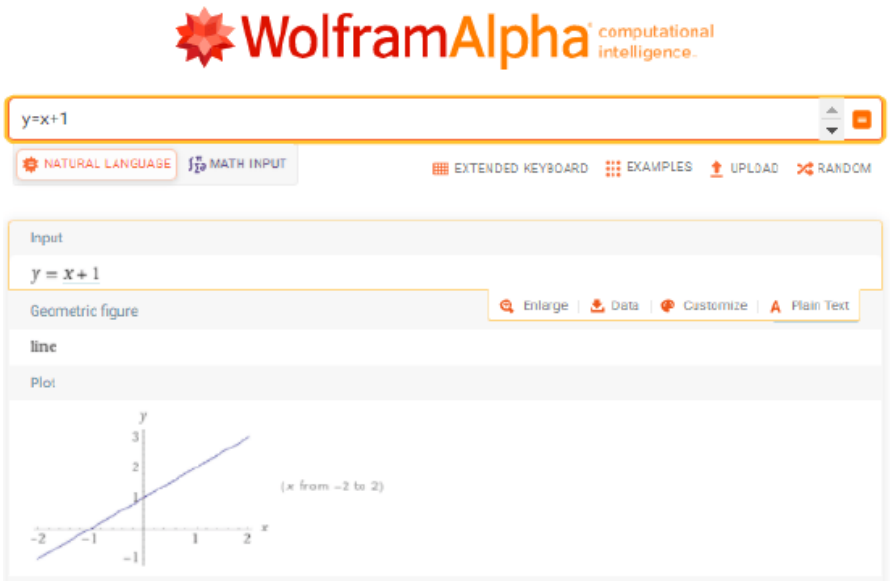
Graphing with Technology

Luckily, when it comes to graphing, we can trust technology with performing this task. Instead of plugging in the values and plotting by points, let's discuss how technology can be used.

In Desmos Graphing Calculator, enter the equation then press Enter. For example, graphing $y = x + 1$ from Example $1.4.3$ is shown below:



In WolframAlpha, enter the equation then press Enter. For example, graphing $y = x + 1$ from Example $1.4.3$ is shown below:



Answers

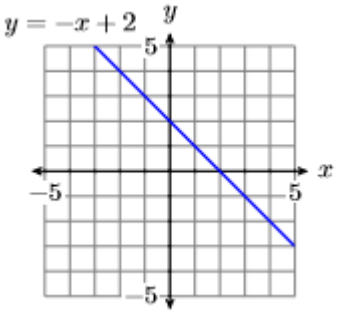
Exercise 1.4.1

$(3, -2)$

Exercise 1.4.2

$(-1, 3)$

Exercise 1.4.3



This page titled 1.4: Graphing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by David Arnold.

- 3.1: Graphing Equations by Hand by David Arnold is licensed CC BY-NC-ND 3.0. Original source: https://math.libretexts.org/@api/deki/files/80188/ElementaryAlgebra.pdf.

2: FINANCE

Close message

Learn more

Welcome to your new guide!

Click on "New" to build out your guide.

- Sections

2.1: Budgeting

2.2: Simple Interest

2.3: Compound Interest

2.4: Installments

This page titled 2: Finance is shared under a license and was authored, remixed, and/or curated by LibreTexts.

2.1: Budgeting

Learning Objectives

In this section, you will learn to:

- Use terms associated with finances (expenses, bills, income, budget, taxes, checking account, savings account etc.)
- Understand the connection between investing and borrowing

So that they can produce a budget with a net positive balance.

Throughout this chapter, we will discuss the definitions and formulas associated with finances. Many of these definitions and formulas are covered in all types of mathematics courses at the college level; however, understanding the mathematics behind these concepts and how they apply to your everyday life are two very different ideas. As you go through this section, and this chapter for that matter, consider how these ideas apply to your own life. Have you ever made a budget? What comes to mind when you think of the differences between income and expenses? When/where do you see yourself investing money? What about borrowing? Review the following definitions and try to name examples of each in your own life.

Definitions

- **Income:** This is money received, usually on a regular basis. Income can be from work or through investments.
- **Expense:** This is money spent on something. Some expenses may be reoccurring, others may be one-time expenses.
- **Budget:** An estimate of expected income and expenses for a set period of time.
- **Taxes:** A obligatory contribution to state and federal revenue, imposed by the government on workers' income and business profits, or added to the cost of some goods, services, and transactions.

It is vital that your income (after taxes) meets, and ideally exceeds, your expenses; in other words, you should try to create a budget with a net positive balance. You can do this is in any time period you would like (e.g Bi-Weekly, Monthly, or Yearly). In the case of budgeting, planning around your income schedule is helpful; having a weekly, or biweekly, budget with a net positive balance will grow monthly and yearly.

Budgeting

In order to create a budget, you should list your income and expenses. Although you can do this in any way that you would like, we'll recommend that you use technology (Google Sheets or Microsoft Excel) to help.

1. Make a list of your income (after taxes).
2. Identify your reoccurring expenses. Be sure to include everything you can here! (Bills, Food, Subscriptions, etc.)
3. Determine whether your income exceeds your expenses.
4. If your income does not exceed your expenses, determine whether there are any expenses you can cut back on; if not, determine how you can increase your income.

It may not be easy to do this process on the first try! Sometimes it takes a couple of months of tracking to nail your budget down. With your Office 365 account, you can use this Personal Monthly Budget spreadsheet to enter your projected budget; then at the end of the month, you can come back and enter your actual budget to compare the difference! This information should be used to plan for future months and examine your spending habits.

There are many budgeting guidelines that can be found online, but sources like Nerdwallet offer a simple budgeting framework called the 50/30/20 rule where 50% of your monthly income goes directly to your needs, 30% of your income goes to your wants, and the last 20% of your income goes to savings and debt repayment; note the savings and debt repayment category can be any payments towards loans that would be considered additional payments as well as building an emergency fund/setting aside money for retirement.

Now that you have an idea of what your finances look like, let's review a couple more financial tools that we can use for our benefit.

Definitions

- **Checking Account:** An account at a bank that allows checks to be drawn by the depositor. Since this account is more fluid, no interest is given on money kept here. Online payment systems like Venmo and PayPal aren't technically checking accounts, but they do operate in a similar manner.
- **Savings Account:** An account at a bank that earns interest.
- **Investments:** Money spent with an expectation of profit or material result.
- **Loans:** A sum of money that is borrowed and expected to be paid back (usually with interest).

Ideally, you've got some leftover money each day, week, month, or year. If so, what will you be doing with it? Putting it into the bank? Investing it somewhere? If not, how will you be dealing with the negative balance? Loans may be an option, but as we'll explore this chapter, interest (simple or compound) can increase the severity of the situation!

Net Positive Balance

In order for your budget to have a net positive balance, your income must exceed your expenses. Taking the sum of your expenses will help you determine the minimum amount of income you need to make a net positive balance.

Example 2.1.1

Given Joseph's expenses below, determine the minimum income needed in order for Joseph to have a net positive budget.

Item	
Food	
Housing	
Transportation	
Entertainment	
Miscellaneous	
Gas	
Phone	
Electricity	
Utilities	

Solution

Adding up Joseph's expenses for Food, Housing, Transportation, Entertainment, Miscellaneous, Gas, Phone, Electricity, and Utilities, we get

$$\$400 + \$600 + \$300 + \$80 + \$310 + \$120 + \$70 + \$45 + \$85 = \$2010$$

This means Joseph should plan to have an income that exceeds $\$2010$ if they want to have a budget with a net positive balance.

You try it 2.1.1

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

View this Interactive Content online

Budgeting for Specific Expenses

Budgeting is necessary at a micro and macro level. Once you've established a monthly budget, you'll need to adjust it for various life events. How does family planning play a role in your budget? What if you're planning a trip or vacation? We can identify how much we need to budget for such things, by identifying as much of the costs as possible.

Example 2.1.1

A couple is getting ready to drive across the USA from Philadelphia, PA to Los Angeles, CA in five days. The distance is approximately $3,000$ miles. The average price of gasoline is about $\$2.00$ per gallon. They plan to stay in a hotel for four nights and budget for the cost of meals to be about $\$50$ per day per person.

1. If their van averages 25 mpg, estimate the total cost of the gas for the trip. Round to the nearest whole number.
2. b) If the average cost of a hotel stay per night is about $\$85$ plus 18 hotel tax, determine the total cost for hotels for their trip. Round answer to the nearest whole number.
3. How much should they budget for the total cost of meals for their trip?

Solution

1. Since the trip is approximately 3000 miles and the van gets an average of 25 mpg, the couple should expect to use $\frac{3000}{25} = 120$ gallons of gas. If the cost of gas is $\$2.00$ the total expected cost of gas for the trip is $120 \cdot 2$=$\$240$.
2. Now, if the average cost of the hotel stay per night is about $\$85$ plus 18 hotel tax, we can say the expected cost of the hotels for 4 nights (before tax) is $4 \cdot 85 = \$340$. Then, the 18 tax on the 4 nights is $340 \cdot 0.18 = \$61$. Adding these totals together we get that the total expected cost is $340 + 61 = \$401$. Note: The alternative way to find this value would be to multiply the 4-night total times 1.18 (the total stay with 18% tax included) so that $340 \cdot 1.18 = \$401$.
3. Finally, since the couple will be staying in a hotel for 4 nights, the trip will take 5 days. With the cost of meals being about $\$50$ per day per person, the couple must budget $\$100$ total for the two of them each day. Multiplying 5 days times $\$100$ for each day gives us a total of $\$500$ to be budgeted for meals.

This means that the couple can expect the total trip to cost about $\$240 + \$401 + \$500 = \$1,141$.

You try it 2.1.1

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

View this Interactive Content online

This page titled 2.1: Budgeting is shared under a license and was authored, remixed, and/or curated by LibreTexts.

- 5.1: Basic Budgeting by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.

2.2: Simple Interest

Learning Objectives

In this section, you will learn to:

- Use the simple interest formula to calculate the future value of a lump sum

So that they can analyze financial scenarios to make informed decisions.

Discussing interest starts with the **principal**, or the amount your account starts with. This could be a starting investment or the starting amount of a loan. Interest, in its most simple form, is calculated as a percent 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 just be 5% of 100: $\$100(0.05) = \$5$. The total amount you would repay would be $105, the original principal plus the interest.

Simple One-time Interest

$$I = P_0 r \tag{2.2.1}$$

$$A = P_0 + I = P_0 + P_0 r = P_0(1 + r) \tag{2.2.2}$$

where

- I is the interest
- A is the end amount: principal plus interest
- P_0 is the principal (starting amount)
- r is the interest rate (in decimal form. Example: $5\% = 0.05$)

Example 2.2.1

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

Solution

$P_0 = \$300$	the principal
$r = 0.03$	3% rate
$I = \$300(0.03) = \$9.$	You will earn $9 interest.

You try it 2.2.1

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

View this Interactive Content online

One-time simple interest is only common for extremely short-term loans. For longer-term loans, it is common for interest to be paid on a daily, monthly, quarterly, or annual basis. In that case, interest would be earned regularly. For example, bonds are essentially a loan made to the bond issuer (a company or government) by you, the bondholder. In return for the loan, the issuer agrees to pay interest, often annually. Bonds have a maturity date, at which time the issuer pays back the original bond value.

Example 2.2.2

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

Solution

Each year, you would earn 5% interest: $\$1000(0.05) = \$50$ in interest. So over the course of five years, you would earn a total of $250 in interest. When the bond matures, you would receive back the $1,000 you originally paid, leaving you with a total of $1,250.

We can generalize this idea of simple interest over time.

Simple Interest over Time

$I = P_0rt$

$A = P_0 + I = P_0 + P_0rt = P_0(1 + rt)$

where

- I is the interest
- A is the end amount: principal plus interest
- P_0 is the principal (starting amount)
- r is the interest rate in decimal form
- t is time

The units of measurement (years, months, etc.) for the time should match the time period for the interest rate.

APR – Annual Percentage Rate

Interest rates are usually given 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 up.

For example, a 6% APR paid monthly would be divided into twelve 0.5% payments.
A 4% annual rate paid quarterly would be divided into four 1% payments.

Example 2.2.3: Treasury Notes

Treasury Notes (T-notes) are bonds issued by the federal government to cover its expenses. Suppose you obtain a $1,000 T-note with a 4% annual rate, paid semi-annually, with a maturity in 4 years. How much interest will you earn?

Solution

Since interest is being paid semi-annually (twice a year), the 4% interest will be divided into two 2% payments.

$P_0 = \$1000$	the principal
$r = 0.02$	2% rate
$t = 8$	4 years $= 8$ half-years
$I = \$1000(0.02)(8) = \$160.$	You will earn $160 interest total over the four years.

You try it 2.2.2

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

View this Interactive Content online

This page titled 2.2: Simple Interest is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Lippman (The OpenTextBookStore).

- 9.2: Simple Interest by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

2.3: Compound Interest

Learning Objectives

In this section, you will learn to:

- Use the compound interest formula to calculate the future value of a lump sum

So that they can analyze financial scenarios to make informed decisions.

With simple interest, we were assuming that we pocketed the interest when we received it. In a standard bank account, any interest we earn is automatically added to our balance, and we earn interest on that interest in future years. This reinvestment of interest is called **compounding**.

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 $\frac{3\%}{12} = 0.25\%$ per month.

In the first month,

$P_0 = \$1000$

$r = 0.0025(0.25\%)$

$I = \$1000(0.0025) = \$2.50$

$A = \$1000 + \$2.50 = \$1002.50$

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

In the second month,

$P_0 = \$1002.50$

$I = \$1002.50(0.0025) = \$2.51$ (rounded)

$A = \$1000 + \$2.50 = \$1002.50$

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** of interest gives us.

Calculating out a few more months:



To find an equation to represent this, if P_m represents the amount of money after m months, then we could write the recursive equation:

$P_0 = \$1000$

$P_m = (1 + 0.0025)P_{m-1}$

You probably recognize this as the recursive form of exponential growth. If not, we could go through the steps to build an explicit equation for the growth:

$P_0 = \$1000$

$P_1 = 1.0025P_0 = 1.0025(1000)$

$P_2 = 1.0025P_1 = 1.0025(1.0025(1000)) = 1.0025^2(1000)$

$P_3 = 1.0025P_2 = 1.0025\left(1.0025^2(1000)\right) = 1.0025^3(1000)$

$P_4 = 1.0025P_3 = 1.0025\left(1.0025^3(1000)\right) = 1.0025^4(1000)$

Observing a pattern, we could conclude

$P_m = (1.0025)^m(\$1000)$

Notice that the \$1000 in the equation was P_0, the starting amount. We found 1.0025 by adding one to the growth rate divided by 12, since we were compounding 12 times per year.

Generalizing our result, we could write

$P_m = P_0\left(1 + \frac{r}{k}\right)^m$

In this formula:

m is the number of compounding periods (months in our example)

r is the annual interest rate

k is the number of compounds per year.

While this formula works fine, it is more common to use a formula that involves the number of years, rather than the number of compounding periods. If N is the number of years, then $m = Nk$. Making this change gives us the standard formula for compound interest.

Compound Interest

$P_N = P_0\left(1+\frac{r}{k}\right)^{Nk}$

P_N is the balance in the account after N years.

P_0 is the starting balance of the account (also called an initial deposit, or principal)

r is the annual interest rate in decimal form

k is the number of compounding periods in one year.

If the compounding is done annually (once a year), $k = 1$.

If the compounding is done quarterly, $k = 4$.

If the compounding is done monthly, $k = 12$.

If the compounding is done daily, $k = 365$.

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.

Example 2.3.1

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?

Solution

In this example,

$P_0 = \$3000$	the initial deposit
$r = 0.06$	6% annual rate
$k = 12$	12 months in 1 year
$N = 20$	since we're looking for how much we'll have after 20 years

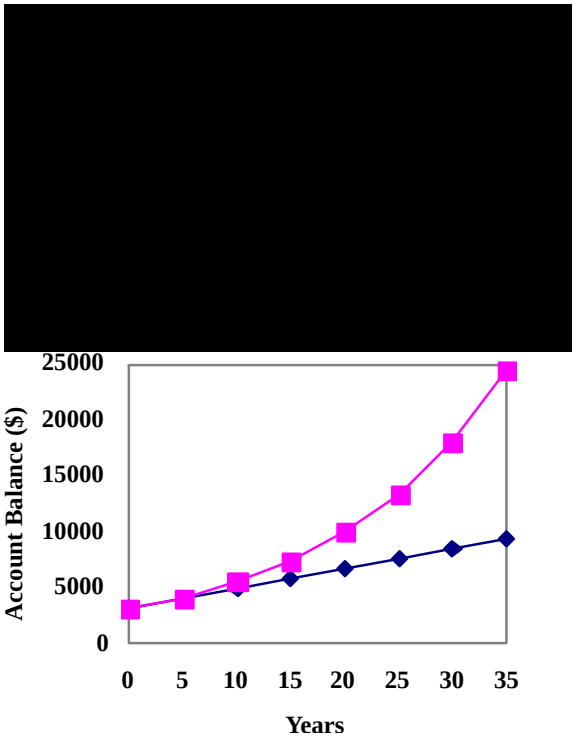
So $P_{20} = 3000\left(1+\frac{0.06}{12}\right)^{20\times 12} = \$9930.61$ (round your answer to the nearest penny)

You try it 2.3.1

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

View this Interactive Content online

Let us compare the amount of money earned from compounding against the amount you would earn from simple interest



As you can see, 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.

Evaluating exponents on the calculator

When we need to calculate something like 5^3 it is easy enough to just multiply $5 \cdot 5 \cdot 5 = 125$. But when we need to calculate something like 1.005^{240}, it would be very tedious to calculate this by multiplying 1.005 by itself 240 times! So to make things easier, we can harness the power of our scientific calculators.

Most scientific calculators have a button for exponents. It is typically either labeled like:

$[\wedge]$, $[y^x]$, or $[x^y]$

To evaluate 1.005^{240} we'd type 1.005 $[\wedge]$ 240, or 1.005 $[y^x]$ 240. Try it out - you should get something around 3.3102044758.

Example 2.3.2

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?

Solution

We're looking for P_0.

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

In this case, we're going to have to set up the equation, and solve for P_0.

$40000 = P_0\left(1+\frac{0.04}{4}\right)^{18\times 4}$

$40000 = P_0(2.0471)$

$P_0 = \frac{40000}{2.0471} = \$19539.84$

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

You try it 2.3.2

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

View this Interactive Content online

Rounding

It is important to be very careful about rounding when calculating things with exponents. In general, you want to keep as many decimals during calculations as you can. Be sure to **keep at least 3 significant digits** (numbers after any leading zeros). Rounding 0.00012345 to 0.000123 will usually give you a "close enough" answer, but keeping more digits is always better.

Example 2.3.3

To see why not over-rounding is so important, suppose you were investing $1000 at 5% interest compounded monthly for 30 years.

Solution

$P_0 = \$1000$	the initial deposit
$r = 0.05$	5%
$k = 12$	12 months in 1 year
$N = 30$	since we're looking for the amount after 30 years

If we first compute $\frac{r}{k}$, we find $\frac{0.05}{12} = 0.00416666666667$

Here is the effect of rounding this to different values:



If you're working in a bank, of course you wouldn't round at all. For our purposes, the answer we got by rounding to 0.00417, three significant digits, is close enough - $5 off of $4500 isn't too bad. Certainly keeping that fourth decimal place wouldn't have hurt.

Using your calculator

In many cases, you can avoid rounding completely by how you enter things in your calculator. For example, in the example above, we needed to calculate

$P_{30} = 1000\left(1 + \frac{0.05}{12}\right)^{12\times 30}$

We can quickly calculate $12 \times 30 = 360$, giving $P_{30} = 1000\left(1 + \frac{0.05}{12}\right)^{360}$.

Now we can use the calculator.



Using your calculator continued

The previous steps were assuming you have a "one operation at a time" calculator; a more advanced calculator will often allow you to type in the entire expression to be evaluated. If you have a calculator like this, you will probably just need to enter:

1000 [$\times$] (1 [$+$] 0.05 [$\div$] 12) [y^x] 360 [$=$].

Annual Percent Yield (APY)

Often when reviewing interest rates of financial products, you hear about the Annual Percent Rate and Annual Percent Yield (APR and APY). What is the difference between these two? **Annual Percent Yield** is the effective interest rate at the end of the year when all of the compoundings have occurred. Notice that in our opening example, we started with \$1000 invested at 3% interest compounded monthly; at the end of the year we end up with \$1030.42 which means that there was a simple yield of 3.042% interest on the original \$1000 investment. 3% interest compounded monthly for a full year results in an annual percent yield of 3.042%.

APY Calculations

The Annual Percent Yield of a given financial product is given by

$$\text{APY}=(1 + \tfrac{r}{n})^n - 1$$

where $r =$ the annual percentage rate and $n =$ the number of compoundings per year.

Notice that applying this formula to 3% annual interest compounded monthly, we get $(1 + \frac{.03}{12})^{12} - 1 = 0.03041595691$ which is 3.042% when rounded.

Example 2.3.4

A bank features a savings account that has an annual percentage rate of $r = 3$. Find the Annual Percentage Yield (APY) for the savings account if the interest is compounded according to the following schedules:

1. Annually

2. Semi-annually
3. Quarterly
4. Monthly

Solution

The compound interest formula is

$$A = P(1 + \frac{r}{n})^{nt}$$

but to find the annual percenaget yield, we can use the formula:

$$APY = (1 + \frac{r}{n})^{n} - 1$$

Depending on the frequency of the compoundings, n may vary:

- $n = 1$ if the compoundings are Annual;
- $n = 2$ if the compoundings are Semiannual;
- $n = 4$ if the compoundings are Quarterly;
- $n = 12$ if the compoundings are Monthly.

Using the $r = 0.03$ and the various compounding periods, we have

1. Annual: $APY = (1 + \frac{0.03}{1})^{1} - 1 = 0.03 = 3$
2. Semiannual: $APY = (1 + \frac{0.03}{2})^{2} - 1 = 0.030225 \approx 3.022$
3. Quarterly: $APY = (1 + \frac{0.03}{4})^{4} - 1 = 0.030339190664063 \approx 3.034$
4. Monthly: $APY = (1 + \frac{0.03}{12})^{12} - 1 = 0.030415956913507 \approx 3.042$

You try it 2.3.3

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

View this Interactive Content online

This page titled 2.3: Compound Interest is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Lippman (The OpenTextBookStore).

- 9.3: Compound Interest by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

2.4: Installments

Learning Objectives

In this section, you will learn to:

- Use technology to
  - Determine the payment schedule of a payout annuity (loan);
  - Identify the amount of interest paid on each payment;
  - Identify the amount paid on the balance of the payout annuity (loan);
  - Identify the remaining balance of the payout annuity (loan);
  - Identify the loan amount that is affordable based on a fixed budget amount;

So that they can analyze financial scenarios to make informed decisions.

For most of us, we aren't able to put a large sum of money in the bank today. Instead, we save for the future by depositing a smaller amount of money from each paycheck into the bank. This idea is called a **Savings Annuity**. Most retirement plans like 401k plans or IRA plans are examples of savings annuities.

There is also a variation called a **Payout Annuity**. With a payout annuity, you start with money in the account, and pull money out of the account on a regular basis. Any remaining money in the account earns interest. After a fixed amount of time, the account will end up empty. Payout annuities are typically used after retirement. Perhaps you have saved $500,000 for retirement, and want to take money out of the account each month to live on. You want the money to last you 20 years. This is a payout annuity.

Annuities

Annuities assume that money is transferred on a regular schedule (every month, year, quarter, etc.) and let the amount sit there earning interest.

Compound interest assumes that you put money in the account once and let it sit there earning interest.

Compound interest: One deposit

Savings Annuity: Many deposits.

Payout Annuity: Many withdrawals

In this section, you will learn about conventional loans (also called amortized loans or installment loans). Examples include auto 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.

Loans are just payout annuities! To see why, imagine that you had $10,000 invested at a bank, and started taking out payments while earning interest as part of a payout annuity, and after 5 years your balance was zero. Flip that around, and imagine that you are acting as the bank, and a car lender is acting as you. The car lender invests $10,000 in you. Since you're acting as the bank, you pay interest. The car lender takes payments until the balance is zero.

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 have to pay from the sale.

To determine the remaining loan balance after some number of years, we first need to know the loan payments, if we don't already know them. 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.

Amortization Calculators

Amortization Calculators use the interested rates, compounding periods, time, and principal amount to determine the amount of interest and principal paid, as well as the remaining balance.

For example, suppose you had a $12,000 loan with 2.9% interest (compounded monthly) and you paid $1000 each month, you would owe an additional $191.87 after one year. See the payment schedule below:

Month	Starting Balance	You Paid	Interest	Principal	Ending Balance
1	$12,000.00	$1,000.00	$29.00	$971.00	$11,029.00
2	$11,029.00	$1,000.00	$26.65	$973.35	$10,055.65
3	$10,055.65	$1,000.00	$24.30	$975.70	$9,079.95
4	$9,079.95	$1,000.00	$21.94	$978.06	$8,101.90
5	$8,101.90	$1,000.00	$19.58	$980.42	$7,121.48
6	$7,121.48	$1,000.00	$17.21	$982.79	$6,138.69
7	$6,138.69	$1,000.00	$14.84	$985.16	$5,153.52
8	$5,153.52	$1,000.00	$12.45	$987.55	$4,165.98
9	$4,165.98	$1,000.00	$10.07	$989.93	$3,176.04
10	$3,176.04	$1,000.00	$7.68	$992.32	$2,183.72
11	$2,183.72	$1,000.00	$5.28	$994.72	$1,189.00
12	$1,189.00	$1,000.00	$2.87	$997.13	$191.87

.

Example 2.4.1

Use this Amortization Calculator (Known Loan Amount) to find the minimum monthly payment on a $20,000 car loan with 2.9% interest over 6 years.

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

View this Interactive Content online

Solution

After entering a loan amount as 20000, the interest rate as .029, and the term in years as 6, you should receive a monthly payment of $302.98

You Try It 2.4.1

Use the calculator linked above to solve the following problem:

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

View this Interactive Content online

This information can be truly eye-opening! The amount of interest paid over the life of these loans increases the price of the object you're buying by much more than expected. Knowing how to calculate the payment on a loan is important, but many times the question is reversed; that is, knowing how much your monthly budget will allow, what loan can you afford? Can you use the Amortization Calculator above to determine the loan amount you can afford with $300 a month payments for three years? What about 6 years? Although it works, guessing and checking the principal loan amount is inefficient.

Example 2.4.2

Use this Amortization Calculator (Known Monthly Payment) to find the loan amount with a $500 monthly payment on a loan with 2.9% interest over 6 years.

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

View this Interactive Content online

Solution

After entering a payment amount as 500, the interest rate as .029, and the term in years as 6, you should receive a maximum loan amount of $33005.52

You Try It 2.4.2

Use the calculator linked above to solve the following problem:

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

View this Interactive Content online

As you consider your financial future (i.e student loans, car loans, or mortgages) be sure to use these tools to help you make informed decisions about what you can afford!

This page titled 2.4: Installments is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Lippman (The OpenTextBookStore).

- 9.4: Annuities by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.
- 9.6: Loans by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.
- 9.7: Remaining Loan Balance by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

3: SET THEORY AND LOGIC



- Topic hierarchy

3.1: The Basics

3.2: Operations with Sets

3.3: Venn Diagrams

3.4: Inductive and Deductive Reasoning

3.5: Using Venn Diagrams to Analyze Arguments

3.6: Common Valid and Invalid Arguments

3.7: Truth Tables

3.8: Analyzing Arguments with Truth Tables

This page titled 3: Set Theory and Logic is shared under a license and was authored, remixed, and/or curated by LibreTexts.

3.1: The Basics

Learning Objectives

Students will be able to:

- Define a set using the roster method and set-builder notation
- Identify the elements of a set
- Compare and contrast the concepts of set, subset, universal set, and the empty set

Definition: Set and Elements

A **set** is a collection of objects or things. The objects or things in the set are called **elements** of the set. A set is commonly denoted by a capital letter and defined by describing the contents, or by listing the elements of the set, enclosed in braces. For example, set A, containing the elements 2, 3, and 4, can be written as A = {2, 3, 4}.

Example 3.1.1

What is the name of the given set below and what are its elements?

C = {1, 6, 2, 5}

Solution

The name of the set is "C" and its elements are 1, 6, 2, and 5.

In the previous example, the set was defined by listing each element of the set enclosed in braces. This method of defining a set is called the **roster method**. In the next example, the set is defined by using a method called **set-builder notation.** Like the roster method, the elements are enclosed in braces. However, instead of listing the individual elements separated by commas as in the roster method, the set-builder notation utilizes a description of the elements. This is accomplished by first assigning a variable, usually a lower case letter, then describing the variable. Set-builder notation is often used for sets in which the number of elements is too large or impossible to write using the roster method.

Example 3.1.2

What is the name of the given set below and what are its elements? Also, write the set in roster notation.

G = {x | x is a counting number less than 5}

Solution

The set of symbols, "G = {x | x is a counting number less than 5}" can be translated into English as, "The set G is the set of all x, such that x is a counting number less than 5". The name of the set is "G" and its elements are 1, 2, 3, and 4. In roster notation: G = {1, 2, 3, 4}

Let's suppose B = {y | y is a food that is tasty}. The problem with this set is that depending on personal preference, the elements of B may be different for different people. I did not think onions are "tasty" so I would not include them in B. However, my son thinks onions are "tasty" and therefore, he would incude onions in B. So, we must be sure to clearly specify a set so that in every situation, the elements of the set would be the same. This is the idea behind **"well-defined".**

Definition: Well-Defined

A set is **well-defined** if given any object, it is clear that the object is or is not an element of the set.

Example 3.1.3

Determine if the given sets are well-defined

a) B = {y | y is a food that is tasty}

b) W = {w | w is a Grammy Award-winning actress prior to 2021}

c) M = { m | m is a fast car}

d) V = { v | v is a past Vice President of The United States}

Solutions

a) B is not well-defined since different people would consider different foods tasty and therefore B may contain different elements for different people.

b) W is a well-defined set because an actress was either a Grammy Award-winner or not a Grammy Award-winner prior to 2021.

c) M is not well-defined since the concept of what is "fast" for a car could vary. Is a 1968 Volkswagen Beetle fast? What is fast?

d) V is a bit tricky because depending on the time frame, a person may or may not be included in the set. We must consider that at the time we are making the determination, the list of past Vice Presidents of The United States may be determined, therefore the set is well-defined even though in the future, the elements of the set will most likely change.

Let A = { x | x is a vowel of the alphabet}. It should be clear that this set A is not well-defined because the description of the elements does not specify which alphabet the vowels are coming from (other languages have vowels) despite the fact that the set was described in English. We can use the idea of a Universal Set to help restrict the context in which the elements are considered.

Definition: Universal Set

The **universal set,** usually denoted by the capital letter U, is a set that establishes context. It contains all the elements within the specific context that is being examined.

Example 3.1.4

Consider the set G = {x | x is a rational number less than 2} and the universe set U = {1, 2, 3, 4}. Use roster notation to rewrite set G.

Solution

G = {1}. Without the universe established the way it was, it would be quite impossible to write G using the roster notation because there is an infinite number of rational numbers less than 2. Since the universe only contained a few numbers and the number 1 was the only element of the universe that met the description in set G, we have the answer as G = {1}. Please note, changing the universe can result in different answers.

Definition: Empty Set

A set that contains no elements is called the **empty set** and is notated as { } or $\emptyset$.

Example 3.1.5

Write the set W in roster notation when W={x | x is a woman and a president of the United States}

Solution

Now at the time I write this, this set does not have any elements. So W = { } or W = Ø

Definition: Subset

A **subset** of a set A is another set that contains only elements from the set A, but may not contain all the elements of A.

If B is a subset of A, we write $B \subseteq A$.

A **proper subset** is a subset that is not identical to the original set – it contains fewer elements.

If B is a proper subset of A, we write $B \subset A$.

Example 3.1.6

Consider these three sets:

A = the set of all even numbers, B = {2, 4, 6}, and C = {2, 3, 4, 6}

a) $B \subset A$ (B is a proper subset of A) since every element of B is also an element of A (every element of B is an even number).

b) It is also true that $B \subset C$.

c) C is not a subset of A, since C contains an element, 3, that is not contained in A

This page titled 3.1: The Basics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- 3.1: Basics by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf.

3.2: Operations with Sets

Learning Objectives

Students will be able to:

- Determine the union and intersection of two or more sets
- Determine the complement of a set
- Determine the difference between sets

In this section, we will look at building new sets from given sets. To do this, we will define and use several "operations." This has rather practical applications. For example, you and a new roommate decide to have a house party, and you both invite your set of friends. At this party, two sets are being combined, though it might turn out that there are some friends that were in both sets.

Definition: Union

The **union** of two sets contains all the elements contained in either set (or both sets). The union is notated $A \bigcup B$.

Example 3.2.1

a) Given the sets *A* = {Daryl, Glenn, Rick} and *B* = {Maggie, Michonne}, find their union.

b) Given the sets *A* = {Daryl, Glenn, Rick, Steve} and *B* = {Maggie, Michonne, Steve}, find their union.

Solution

a) *A* ∪ *B* = {Daryl, Glenn, Maggie, Michonne, Rick}. Notice EACH member of each set is listed in the union.

b) *A* ∪ *B* = {Daryl, Glenn, Maggie, Michonne, Rick, Steve}. Notice again that each member of each set is in the union. Also, notice that "Steve" belongs in both sets but is only listed once in the union. Elements in sets are always only listed once.

Definition: Intersection

The **intersection** of two sets contains only the elements that are in both sets. The intersection is notated $A \bigcap B$.

Example 3.2.2

Given the sets *A* = {Aquaman, Batman, Flash, Superman, Wonder Woman} and *B* = {Batman, Blue Beetle, Booster Gold, Fire, Flash}, find the intersection of A and B.

Solution

A ∩ *B* = {Batman, Flash}. Notice Batman and Flash are members of BOTH sets.

Definition: Complement

The **complement** of a set A contains everything that is *not* in the set A but still in the universal set U. The complement is notated A', or A^c, or sometimes $\sim A$.

Example 3.2.3

Let *U* = {1, 2, 3, 4, 5, 6, 7, 8, 9} and *A* = {1, 3, 5, 7, 9}. Find A'.

Solution

A' = {2, 4, 6, 8}. These are all the elements in U that ARE NOT in A.

Definition: Difference

The **difference** between two sets A and B, denoted A\B is the set of all the elements in A but not in B. Using complement notation, A\B = $A \bigcap B^c$

Example 3.2.4

Let A = {T-Mobile, Apple, Verizon} and B = {Spectrum, Apple, Duo}. Find A\B.

Solution

A\B = {T-Moblie, Verizon} since these are the only elements in A but not in B. The element "Apple" is not in the difference because even though it is in A, it is also in B.

We may perform multiple operations on sets. The next few examples illustrate how to do that.

Example 3.2.5

Suppose H = {cat, dog, rabbit, mouse}, F = {dog, cow, duck, pig, rabbit} W = {duck, rabbit, deer, frog, mouse}

a. Find $(H \bigcap F) \bigcup W$
b. Find $H \bigcap (F \bigcup W)$
c. Find $(H \bigcap F)^c \bigcap W$

Solution

a) We start with the intersection: $H \bigcap F$ = {dog, rabbit}

Now we union that result with W: $(H \bigcap F) \bigcup W$ = {dog, rabbit) ⋄ {duck, rabbit, deer, frog, mouse} = {dog, duck, rabbit, deer, frog, mouse}

b) We start with the union: $F \bigcup W$ = {dog, cow, rabbit, duck, pig, deer, frog, mouse}

Now we intersect that result with H: $H \bigcap (F \bigcup W)$ = H = {cat, dog, rabbit, mouse} ⋄ {dog, cow, rabbit, duck, pig, deer, frog, mouse} = {dog, rabbit, mouse}

c) We start with the intersection: $H \bigcap F$ = {dog, rabbit}

Now we want to find the elements of W that are not in $H \bigcap F$.

$(H \bigcap F)^c \bigcap W$ = {duck, deer, frog, mouse}

This page titled 3.2: Operations with Sets is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- 3.2: Union, Intersection, and Complement by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf.

3.3: Venn Diagrams

Learning Objectives

Students will be able to:

- Use Venn Diagrams to illustrate operations with sets

Definition: Venn Diagram

A Venn diagram represents each set by a circle, usually drawn inside of a containing box representing the universal set. Overlapping areas indicate elements common to both sets.

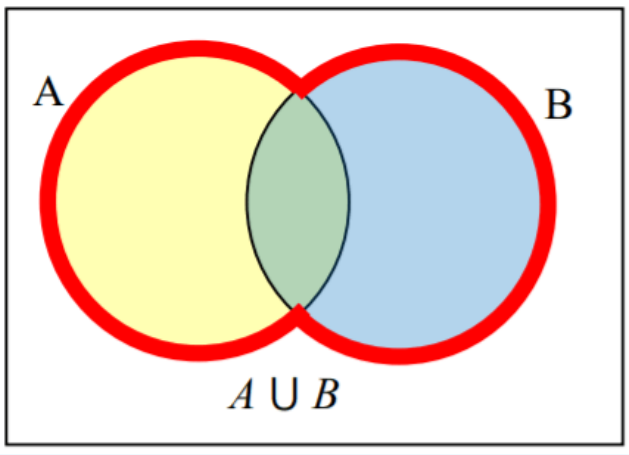
Basic Venn diagrams can illustrate the interaction of two or three sets.

Example 3.3.1

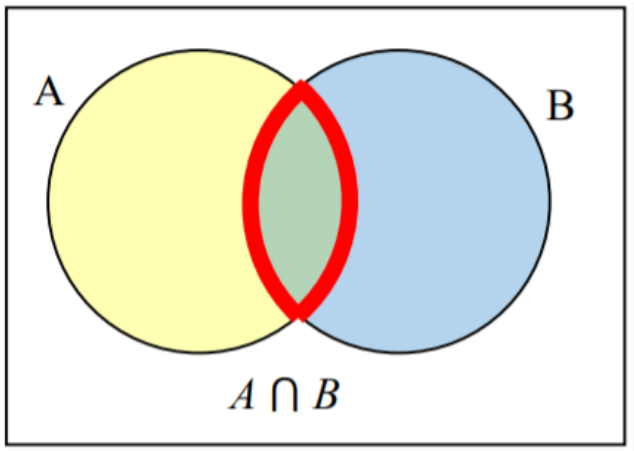
Create Venn diagrams to illustrate $A \bigcup B$, $A \bigcap B$, and $A^c \bigcap B$.

Solution

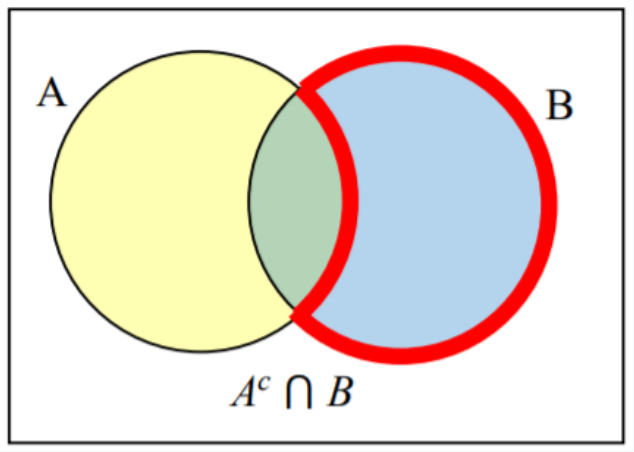
$A \bigcup B$ contains all elements in *either* set.



$A \bigcap B$ contains only those elements in both sets – in the overlap of the circles.



A^c will contain all elements not in the set A. $A^c \bigcap B$ will contain the elements in set B that are not in set A.

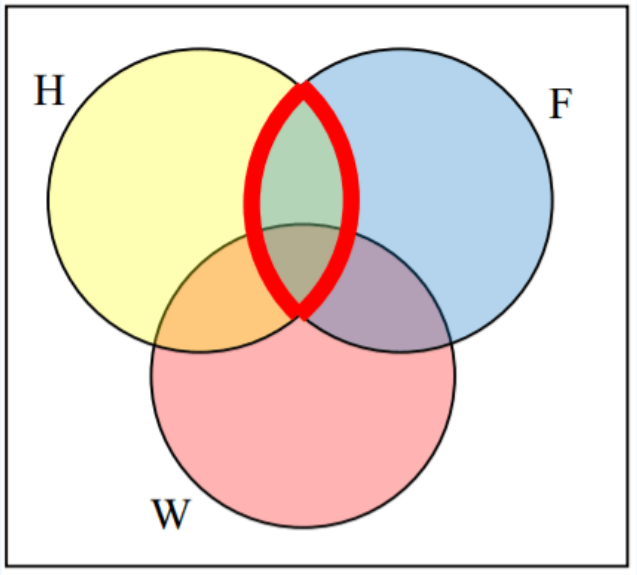


Example 3.3.2

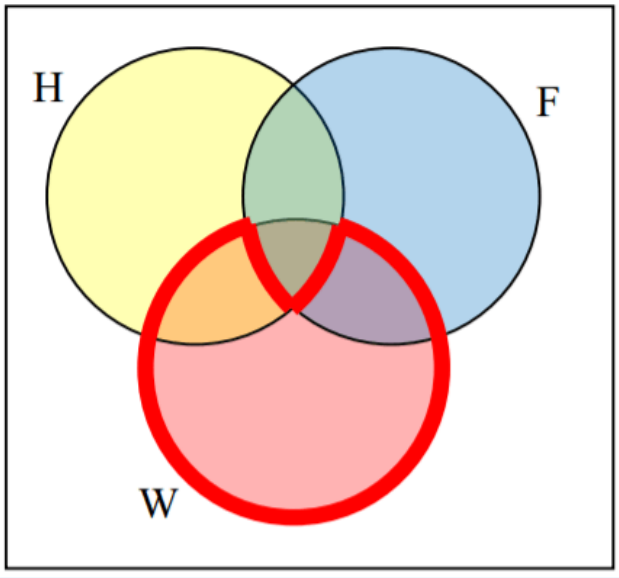
Use a Venn diagram to illustrate $(H \bigcap F)^c \bigcap W$.

Solution

We'll start by identifying everything in the set $H \bigcap F$:

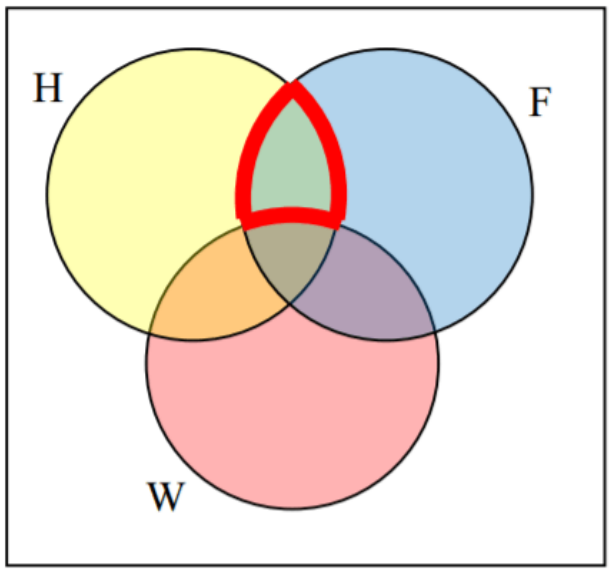


Now, $(H \bigcap F)^c \bigcap W$ will contain everything not in the set identified above that is also in set W.



Example 3.3.3

Create an expression to represent the outlined part of the Venn diagram shown.



Solution

The elements in the outlined set *are* in sets H and F, but are not in set W. So, we could represent this set as $H \bigcap F \bigcap W^c$.

This page titled 3.3: Venn Diagrams is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- 3.3: Venn Diagrams is licensed CC BY-NC-SA 4.0.

3.4: Inductive and Deductive Reasoning

Learning Objectives

Students will be able to

- Identify and utilize deductive and inductive reasoning

Ask somebody who has a job as to why they have a job and there is a good chance they have a reason or multiple reasons. Most likely they will respond by saying that they need the money for their basic necessities. They may even respond that they just want to keep busy or that their parents told them they had to. The point is that there are reasons.

Definition: Reasoning

Reasoning is the act of drawing a *conclusion* from assumed fact(s) called *premise(s)*.

Examples 3.4.1

Identify the premise(s) and conclusion in each case of reasoning:

a) "Martha wants to buy a new smartphone, so she decides to get a job."

b) The traffic app notifies Pedro that the traffic on Interstate 215 North will cause him to arrive at his destination at 3 p.m., an hour later than he expected. The app also shows that Interstate 15 North will allow him to arrive at his destination at 2:30 p.m. Pedro decides to take the Interstate 15 North.

Solutions

a) The premise is that Martha wants to buy a new smartphone and the conclusion is that she decides to get a job.

b) There are two premises in this example. One, that the traffic on Interstate 215 North will cause Pedro to arrive at his destination at 3 p.m, and the other that Interstate 15 North will allow him to arrive at his destination at 2:30 p.m. The conclusion is that Pedro takes the Interstate 15 North.

There are many different forms of reasoning defined by scholars, two of which are defined below.

Definitions: Inductive and Deductive Reasoning

Inductive reasoning: uses a collection of **specific instances** as premises and uses them to propose a **general conclusion**.

Deductive reasoning: uses a collection of **general statements** as premises and uses them to propose a **specific conclusion**.

Notice carefully how both forms of reasoning have both premises and a conclusion. The important difference between these two types is the nature of the premises and conclusion. Applying these definitions to some examples should illuminate the differences and similarities.

Examples 3.4.2

Identify the premises and conclusion of the reasoning below. Identify the type of reasoning used and explain your choice.

a) “When I went to the store last week I forgot my purse, and when I went today I forgot my purse. I always forget my purse when I go to the store”

b) “Every day for the past year, a plane flies over my house at 2 p.m. A plane will fly over my house every day at 2 p.m.”

c) "All electronic devices are useful. My cell phone is an electronic device. Therefore, my cell phone is useful."

d) Spicy food makes me teary. Habanero sauce is spicy food. Habanero sauce makes me teary.

Solution

a) The premises are:

- When I went to the store last week I forgot my purse.
- When I went today I forgot my purse.

The conclusion is:

- I always forget my purse when I go to the store

This is an example of inductive reasoning because the premises are **specific instances,** while the conclusion is **general.**

b) The premise is:

- Every day for the past year, a plane flies over my house at 2 p.m

The conclusion is:

- A plane will fly over my house every day at 2 p.m.

This is an example of inductive reasoning because the premises are **specific instances,** while the conclusion is **general.**

c) The premises are:

- All electronic devices are useful.
- My cell phone is an electronic device.

The conclusion is:

- My cell phone is useful.

d) The premises are:

- Spicy food makes me teary.
- Habanero sauce is spicy food.

The conclusion is:

- Habanero sauce makes me teary.

This is an example of deductive reasoning because the premises are **general statements,** while the conclusion is **specific.**

This page titled 3.4: Inductive and Deductive Reasoning is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- 4.4: Arguments by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf.

3.5: Using Venn Diagrams to Analyze Arguments

Learning Objectives

Students will be able to

- Use Venn diagrams to analyze arguments

When you hear people arguing, what do you think of? Yelling? Screaming? No good, right? Here arguments are placed within the context of logic. So, what is an argument to the logistician? Read on!

Definition: Argument

An **argument** is a collection of premises intended to support or infer a conclusion. Each premise must be either true or false.

The definition above may seem the same as "reasoning" from the prior section (look back there if needed). The key difference is the additional requirement of the premises being true or false. Take the inductive reasoning, "It was hot yesterday, it is hot today, and therefore it will be hot tomorrow." The idea of "hot" is neither true nor false, rather it is an opinion. The statement is not an argument, just inductive reasoning. In an argument, we must be able to assign the values of either true or false to each premise. Once we are sure we have an argument, we can then analyze it.

Definition: Syllogism

A **syllogism** is a specific type of argument that consists of two premises and a conclusion.

Definition: Valid and Invalid Arguments

An argument is considered **valid** if the conclusion of the argument is guaranteed under its given set of premises. Otherwise, the argument is **invalid.**

The concept of validity does not apply to arguments that involve inductive reasoning. Take, for example, the premises, "I knocked down the first domino, and the first domino knocked down the second, then the second knocked down the third" and the conclusion, "the third domino will knock down the fourth". Are we guaranteed that the fourth domino will fall? No. Surely it is the most likely outcome, but cannot be guaranteed. When analyzing inductive arguments, we describe them as either being "strong" or "weak". Certainly, I would consider the "domino" argument "strong".

So, the only types of arguments we will analyze are deductive arguments. There are two common tools to analyze deductive arguments, Venn Diagrams, and Truth Tables. The remainder of this section will be devoted to determining the validity of syllogisms using Venn Diagram. Truth Tables are a more general and quite powerful tool to determine the validity of an argument and will be presented in a subsequent section.

Analyzing Deductive Arguments with Venn Diagrams

To analyze a deductive argument with a Venn diagram:

1) Draw a Venn diagram based on the premises.

2) The argument is invalid if there is a way to draw the diagram that makes the conclusion false

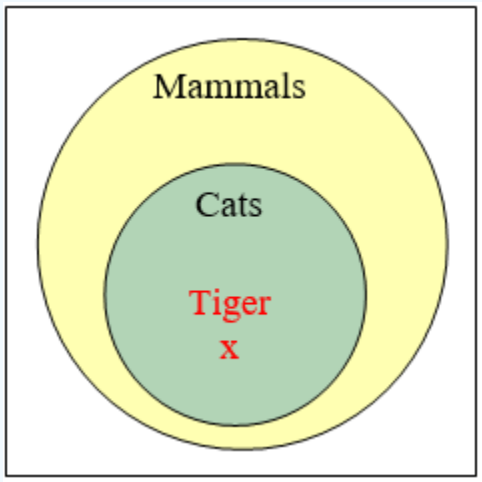
3) The reasoning is valid if the diagram cannot be drawn to make the conclusion false

4) If the premises are insufficient to determine the location of an element or a set mentioned in the conclusion, then the reasoning is invalid.

Example 3.5.1

Consider the statements “All cats are mammals and a tiger is a cat, so a tiger is a mammal.” Identify the premises and conclusion. Is this an argument? Why? and if so is the argument valid? Show with a Venn Diagram.

Solution



The premises are:

- All cats are mammals.
- A tiger is a cat.

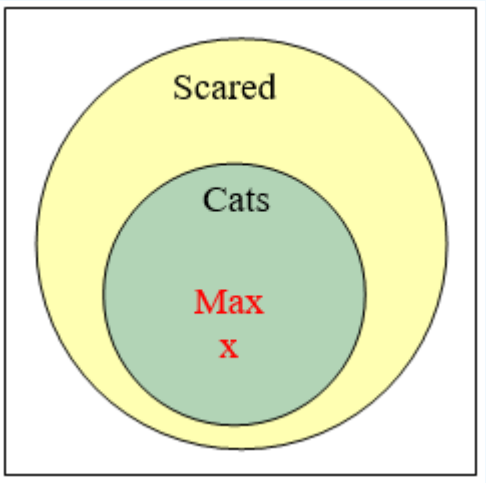
The conclusion is:

- A tiger is a mammal.

This is an argument because the premises are either true or false (they just both happen to be true) and the argument is valid. From the first premise, we draw the set of cats as a subset of the set of mammals. From the second premise, we are told that a tiger is contained within the set of cats. From that, we can see in the Venn diagram that the tiger must also be inside the set of mammals, so there is no way of drawing the diagram to make the conclusion false. This is a valid argument!

A common misunderstanding about the validity of an argument is that the premises must be true. If the premises are all true for a valid argument, then we say the argument is "sound". The next example shows how an argument can still be valid even though there is a false premise. In such circumstances, the argument is called "unsound".

Example 3.5.2



Determine the validity of this argument using a Venn Diagram:

Premise:	All cats are scared of vacuum cleaners.
Premise:	Max is a cat.
Conclusion:	Max is scared of vacuum cleaners.

Solution available at the end of this section.

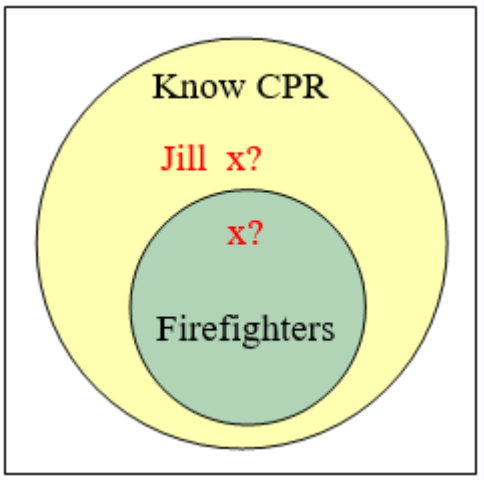
If an argument is invalid, then the argument is unsound. So, both valid and invalid arguments can be "unsound", but only valid arguments can be "sound".

Example 3.5.3

Determine the validity of this argument using a Venn Diagram

Premise:	All firefighters know CPR.
Premise:	Jill knows CPR.
Conclusion:	Jill is a firefighter.

Solution



From the first premise, we know that firefighters all 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 that larger set, but we do not have enough information to know whether she also is a member of the smaller subset that is firefighters.

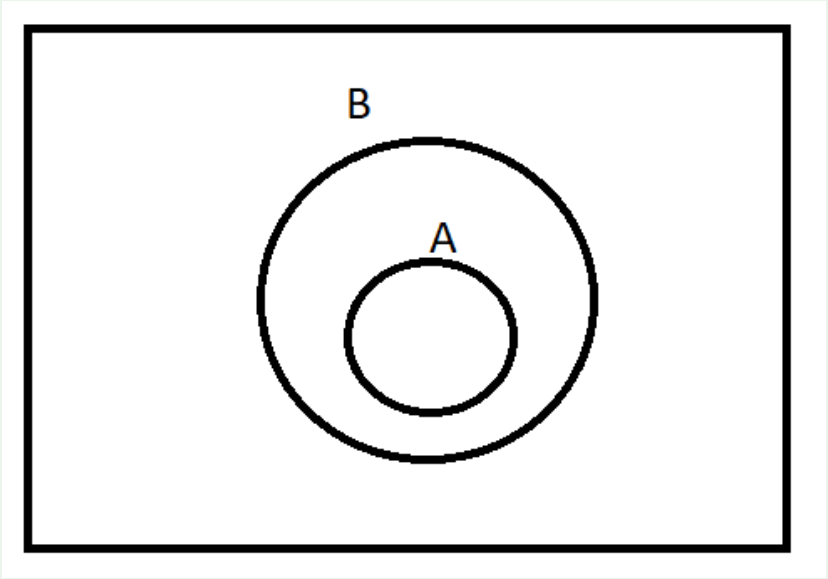
Since the conclusion does not necessarily follow from the premises, this is an invalid argument. It's possible that Jill is a firefighter, but the structure of the argument doesn't allow us to conclude that she definitely is.

It is important to note that whether or not Jill is actually a firefighter is not important in evaluating the validity of the reasoning; we are concerned with whether the premises are enough to prove the conclusion.

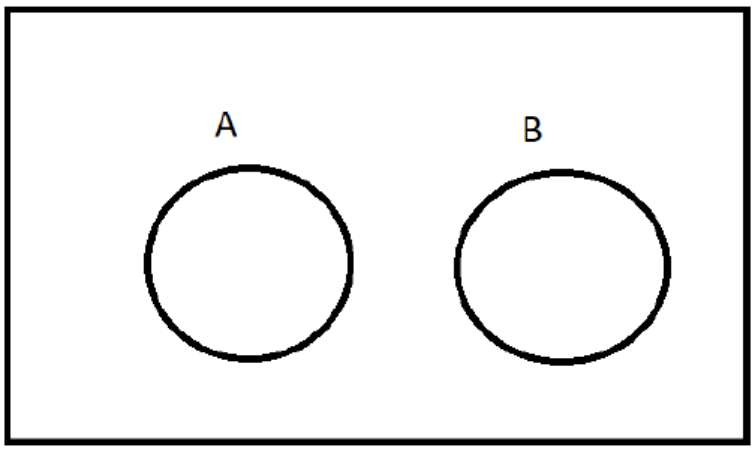
Each example above had a common structure in its premises. "All cats are mammals", "All cats are scared of vacuum cleaners", and "All firefighters know CPR". In general, "All A are B." This commonality is called the Universal Affirmative. Below are four common premise structures that are used in arguments along with the Venn Diagram for each. Making sense of these diagrams should help in creating a Venn Diagram to determine the validity of an argument.

Definitions: Standard Premise Structures

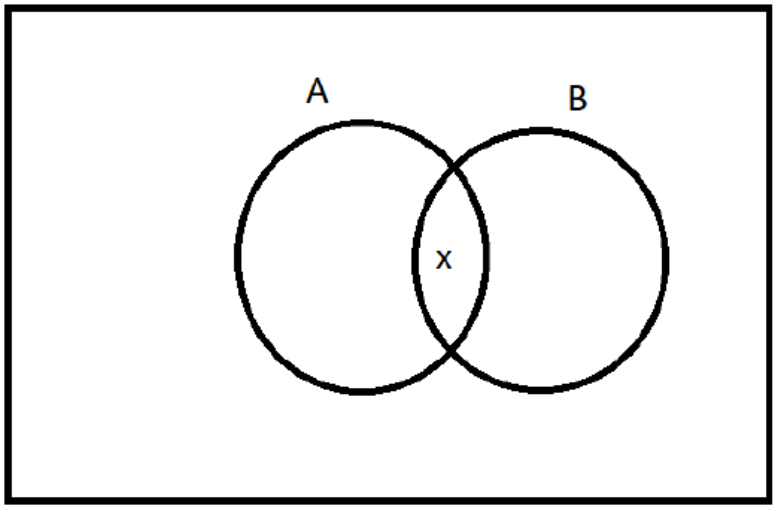
Universal Affirmative: All A are B.



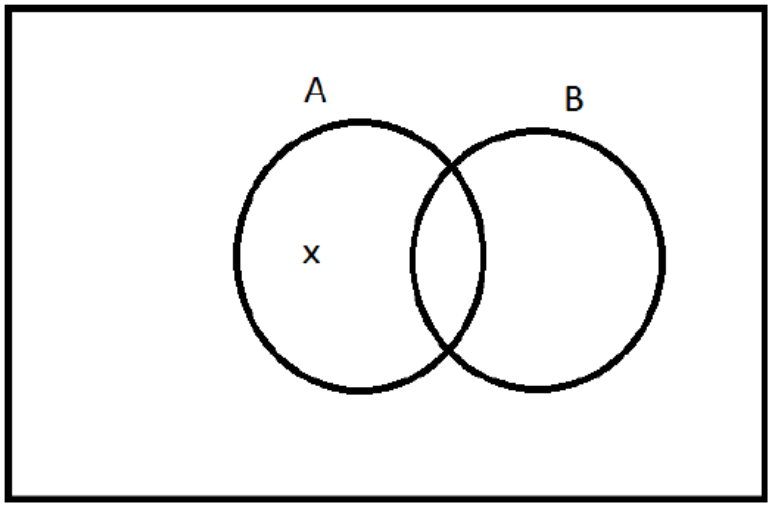
Universal Negative: No A are B.



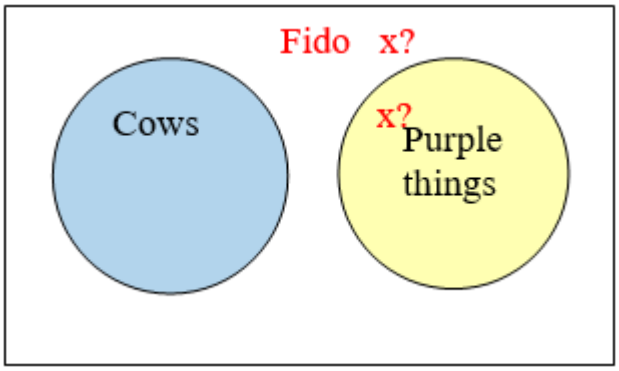
Particular Affirmative: Some A are B. Please note that the "x" signifies that there is at least one A is B



Particular Negative: Some A are not B. The "x" signifies that there is at least one A is not B.



Example 3.5.4



Determine the validity of this argument using a Venn Diagram:

Premise:	No cows are purple.
Premise:	Fido is not a cow.
Conclusion:	Fido is purple.

Solution

The first premise is the universal negative, so there is no overlap between the set of cows and the set of purple things. We know Fido is not in the cow set, but that is not enough to conclude that Fido is in the purple things set. This argument is invalid.

Answers

Example 3.5.2

Valid. Cats are a subset of creatures that are scared by vacuum cleaners. Max is in the set of cats, so he must also be in the set of creatures that are scared by vacuum cleaners. Since the premise "All cats are scared of vacuum cleaners" is false, yet the argument is valid, the argument is unsound.

This page titled 3.5: Using Venn Diagrams to Analyze Arguments is shared under a license and was authored, remixed, and/or curated by LibreTexts.

- 17.9: Evaluating Deductive Arguments with Euler Diagrams by David Lippman is licensed CC BY-SA 3.0. Original source: http://www.opentextbookstore.com/mathinsociety.

3.6: Common Valid and Invalid Arguments

Learning Objectives

Students will be able to:

- Identify common valid and invalid arguments.

It would be difficult to take the time to draw a Venn Diagram to check the validity of every argument you come across. People who argue for a living such as lawyers and judges already know certain argument structures that are always valid, then use them often. Also, they study arguments that are invalid so that they know when they are perhaps being tricked. In this section, you will see valid and invalid argument structures that are quite common. Remember, to be a valid argument, the conclusion must follow from the premises. The premises do not have to be true!

Common Valid Argument Forms

1) **Modus Pones:** All A is/are B.

A.

Therefore, B.

For example: All lawyers study logic.

Zala is a lawyer.

Therefore, Zala studies logic.

2) **Modus Tollens:** All A is/are B.

Not B.

Therefore, not A.

For example: All lawyers study logic.

Patrick has not studied logic.

Therefore, Patrick is not a lawyer.

3) **Hypothetical Syllogism:** All A is/are B.

All B is/are C.

Therefore, All A are C.

For example: All lawyers study logic.

All who study logic are paid well.

Therefore, All lawyers are paid well.

I am sure you will all make valid arguments, especially now that you know some of the basic forms. Be careful however, not everybody has taken a class like this and they will often argue, yet their arguments are invalid. Here are a couple of the most

common invalid arguments.

Common Invalid Argument Forms

1. **Denying the antecedent:** All A is/are B.

Not A.

Therefore, not B.

For example: All lawyers study logic.

Zala is not a lawyer.

Therefore, Zala does not study logic.

2. **Affirming the consequent:** All A is/are B.

B.

Therefore, A

For example: All lawyers study logic.

Patrick studies logic.

Therefore, Patrick is a lawyer

This page titled 3.6: Common Valid and Invalid Arguments is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- 4.5: Logical Fallacies in Common Language by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf.

3.7: Truth Tables

Learning Objectives

Students will be able to

- Translate premises and conclusions into symbols
- Use truth tables to organize simple and compound statements
- Use symbols to express DeMorgan's Laws

Before we focus on truth tables, we're going to introduce some symbols that are commonly used for *and, or,* and *not*.

Symbols

The symbol ⋄ is used for and: A and B is notated A ⋄ B

The symbol ⋄ is used for or: A or B is notated A ⋄ B

The symbol ~ is used for not: not A is notated ~A

You can remember the first two symbols by relating them to the shapes for the union and intersection. A ⋄ B would be the elements that exist in both sets, in A ⋄ B. Likewise, A ⋄ B would be the elements that exist in either set, in A ⋄ B. When we are working with sets, we use the rounded version of the symbols; when we are working with statements, we use the pointy version.

Example 3.7.1

Translate each statement into symbolic notation. Let P represent "I like Pepsi" and let C represent "I like Coke".

a. I like Pepsi or I like Coke.
b. I like Pepsi and I like Coke.
c. I do not like Pepsi.
d. It is not the case that I like Pepsi or Coke.
e. I like Pepsi and I do not like Coke.

Solution

a. P ⋄ C
b. P ⋄ C
c. ~P
d. ~(P ⋄ C)
e. P ⋄ ~C

As you can see, we can use parentheses to organize more complicated statements.

Since premises and conclusions can be quite complicated, a useful tool for organizing everything is a truth table.

Truth table

A table showing what the resulting truth value of a complex statement is for all the possible truth values for the simple statements.

Example 3.7.2

Suppose you're picking out a new couch, and your significant other says "get a sectional or something with a chaise".

Solution

This is a complex statement made of two simpler conditions: "is a sectional", and "has a chaise". For simplicity, let's use S to designate "is a sectional", and C to designate "has a chaise".

A truth table for this situation would look like this:

Table 3.7.1

S	C	S or C
T	T	T
T	F	T
F	T	T
F	F	F

In the table, T is used for true, and F for false. In the first row, if S is true and C is also true, then the complex statement "S or C" is true. This would be a sectional that also has a chaise, which meets our desire. (Remember that *or* in logic is not exclusive; if the couch has both features, it meets the condition.)

In the previous example about the couch, the truth table was really just summarizing what we already know about how the or statement work. The truth tables for the basic *and, or,* and *not* statements are shown below.

Table 3.7.2: *Conjunction*

A	B	A ◈ B
T	T	T
T	F	F
F	T	F
F	F	F

Table 3.7.3: Disjunction

A	B	A ⋄ B
T	T	T
T	F	T
F	T	T
F	F	F

Table 3.7.4: Negation

A	~A
T	F
F	T

Truth tables really become useful when we analyze more complex statements.

Example 3.7.3

When we create the truth table, we need to list all the possible truth-value combinations for A and B. Notice how the first column contains 2 Ts followed by 2 Fs, and the second column alternates T, F, T, F. This pattern ensures that all 4 combinations are considered.

Table 3.7.5

A	B
T	T
T	F
F	T

A	B
F	F

After creating columns with those initial values, we create a third column for the expression ~B. Now we will temporarily ignore the column for A and write the truth values for ~B.

Table 3.7.6

A	B	~B
T	T	F
T	F	T
F	T	F
F	F	T

Next, we can find the truth values of *A* ⧫ *~B*, using the first and third columns.

Table 3.7.7

A	B	~B	A ⧫ ~B
T	T	F	T
T	F	T	T
F	T	F	F
F	F	T	T

The truth table shows that A ⧫ ~B is true in three cases and false in one case. If you're wondering what the point of this is, suppose it is the last day of the baseball season and two teams, who are not playing each other, are competing for the final playoff spot. Anaheim will make the playoffs if it wins its game or if Boston does not win its game. (Anaheim owns the tie-breaker; if both teams win, or if both teams lose, then Anaheim gets the playoff spot.) If A = Anaheim wins its game and B = Boston wins its game, then A ⧫ ~B represents the situation "Anaheim wins its game or Boston does not win its game". The truth table shows us the different scenarios related to Anaheim making the playoffs. In the first row, Anaheim wins its game and Boston wins its game, so it is true that Anaheim makes the playoffs. In the second row, Anaheim wins and Boston does

not win, so it is true that Anaheim makes the playoffs. In the third row, Anaheim does not win its game and Boston wins its game, so it is false that Anaheim makes the playoffs. In the fourth row, Anaheim does not win and Boston does not win, so it is true that Anaheim makes the playoffs.

Example 3.7.4

Create a truth table for the statement A ⊕ ~(B ⊕ C).

Solution

It helps to work from the inside out when creating a truth table, and to create columns in the table for intermediate operations. We start by listing all the possible truth value combinations for A, B, and C. Notice how the first column contains 4 Ts followed by 4 Fs, the second column contains 2 Ts, 2 Fs, then repeats, and the last column alternates T, F, T, F... This pattern ensures that all 8 combinations are considered. After creating columns with those initial values, we create a fourth column for the innermost expression, B ⊕ C. Now we will temporarily ignore the column for A and focus on B and C, writing the truth values for B ⊕ C.

Table 3.7.8

A	B	C	B ⊕ C
T	T	T	T
T	T	F	T
T	F	T	T
T	F	F	F
F	T	T	T
F	T	F	T
F	F	T	T
F	F	F	F

Table 3.7.9

A	B	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

Next, we can find the negation of B ⍰ C, working off the B ⍰ C column we just created. (Ignore the first three columns and simply negate the values in the B ⍰ C column.)

Table 3.7.10

A	B	C	B ⍰ C	~(B ⍰ C)
T	T	T	T	F
T	T	F	T	F
T	F	T	T	F
T	F	F	F	T
F	T	T	T	F

A	B	C	B ⋄ C	~(B ⋄ C)
F	T	F	T	F
F	F	T	T	F
F	F	F	F	T

Finally, we find the values of A and ~(B ⋄ C). (Ignore the second, third, and fourth columns.)

Table 3.7.11

A	B	C	B ⋄ C	~(B ⋄ C)	A ⋄ ~(B ⋄ 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

It turns out that this complex expression is true in only one case: when A is true, B is false, and C is false. To illustrate this situation, 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.

Truth Tables: Conditional, Biconditional

We discussed conditional statements earlier, in which we take an action based on the value of the condition. We are now going to look at another version of a conditional, sometimes called an implication, which states that the second part must logically follow

from the first.

Conditional

A conditional is a logical compound statement in which a statement p, called the antecedent, implies a statement q, called the consequent.

A conditional is written as $p \rightarrow q$ and is translated as "if p, then q".

Example 3.7.5

The English statement "If it is raining, then there are clouds is the sky" is a conditional statement. It makes sense because if the antecedent "it is raining" is true, then the consequent "there are clouds in the sky" must also be true.

Notice that the statement tells us nothing of what to expect if it is not raining; there might be clouds in the sky, or there might not. If the antecedent is false, then the consequent becomes irrelevant.

Example 3.7.6

Suppose you order a team jersey online on Tuesday and want to receive it by Friday so you can wear it to Saturday's game. The website says that if you pay for expedited shipping, you will receive the jersey by Friday. In what situation is the website telling a lie?

Solution

There are four possible outcomes:

1) You pay for expedited shipping and receive the jersey by Friday

2) You pay for expedited shipping and don't receive the jersey by Friday

3) You don't pay for expedited shipping and receive the jersey by Friday

4) You don't pay for expedited shipping and don't receive the jersey by Friday

Only one of these outcomes proves that the website was lying: the second outcome in which you pay for expedited shipping but don't receive the jersey by Friday. The first outcome is exactly what was promised, so there's no problem with that. The third outcome is not a lie because the website never said what would happen if you didn't pay for expedited shipping; maybe the jersey would arrive by Friday whether you paid for expedited shipping or not. The fourth outcome is not a lie because, again, the website didn't make any promises about when the jersey would arrive if you didn't pay for expedited shipping.

It may seem strange that the third outcome in the previous example, in which the first part is false but the second part is true, is not a lie. Remember, though, that if the antecedent is false, we cannot make any judgment about the consequent. The website never said that paying for expedited shipping was the only way to receive the jersey by Friday.

Example 3.7.7

A friend tells you "If you upload that picture to Facebook, you'll lose your job." Under what conditions can you say that your friend was wrong?

Solution

There are four possible outcomes:

1) You upload the picture and lose your job

2) You upload the picture and don't lose your job

3) You don't upload the picture and lose your job

4) You don't upload the picture and don't lose your job

There is only one possible case in which you can say your friend was wrong: the second outcome in which you upload the picture but still keep your job. In the last two cases, your friend didn't say anything about what would happen if you didn't upload the picture, so you can't say that their statement was wrong. Even if you didn't upload the picture and lost your job anyway, your friend never said that you were guaranteed to keep your job if you didn't upload the picture; you might lose your job for missing a shift or punching your boss instead.

In traditional logic, a conditional is considered true as long as there are no cases in which the antecedent is true and the consequent is false.

Truth table for the conditional

Table 3.7.12

p	q	p → q
T	T	T
T	F	F
F	T	T
F	F	T

Again, if the antecedent p is false, we cannot prove that the statement is a lie, so the result of the third and fourth rows is true.

Example 3.7.8

Construct a truth table for the statement (m ⧸ ~p) → r

Solution

We start by constructing a truth table with 8 rows to cover all possible scenarios. Next, we can focus on the antecedent, m ⧸ ~p.

Table 3.7.13

m	p	r
T	T	T
T	T	F
T	F	T

m	p	r
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

Table 3.7.14

m	p	r	~p
T	T	T	F
T	T	F	F
T	F	T	T
T	F	F	T
F	T	T	F
F	T	F	F
F	F	T	T
F	F	F	T

Table 3.7.15

m	p	r	~p	m ⧸ ~p
T	T	T	F	F
T	T	F	F	F
T	F	T	T	T
T	F	F	T	T
F	T	T	F	F
F	T	F	F	F
F	F	T	T	F
F	F	F	T	F

Now we can create a column for the conditional. Because it can be confusing to keep track of all the **Ts** and **Fs**, why don't we copy the column for r to the right of the column for **m ⧸ ~p** ? This makes it a lot easier to read the conditional from left to right.

Table 3.7.16

m	p	r	~p	m ⧸ ~p	r	(m ⧸ ~p) → 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

m	p	r	~p	m ⊽ ~p	r	(m ⊽ ~p) → r
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 antecedent m ⊽ ~p will be true but the consequent false, resulting in an invalid conditional; every other case gives a valid conditional.

If you want a real-life situation that could be modeled by (m ⊽ ~p) → r, consider this: let m = we order meatballs, p = we order pasta, and r = Rob is happy. The statement (m ⊽ ~p) → r is "if we order meatballs and don't order pasta, then Rob is happy". If m is true (we order meatballs), p is false (we don't order pasta), and r is false (Rob is not happy), then the statement is false, because we satisfied the antecedent but Rob did not satisfy the consequent.

For any conditional, there are three related statements, the converse, the inverse, and the contrapositive.

Related Statements

The original conditional is "if p, then q" $p \rightarrow q$

The converse is "if q, then p" $q \rightarrow p$

The inverse is "if not p, then not q" $\sim p \rightarrow \sim q$

The contrapositive is "if not q, then not p" $q \rightarrow \sim p$

Example 3.7.9

Consider again the conditional "If it is raining, then there are clouds in the sky." It seems reasonable to assume that this is true.

The converse would be "If there are clouds in the sky, then it is raining." This is not always true.

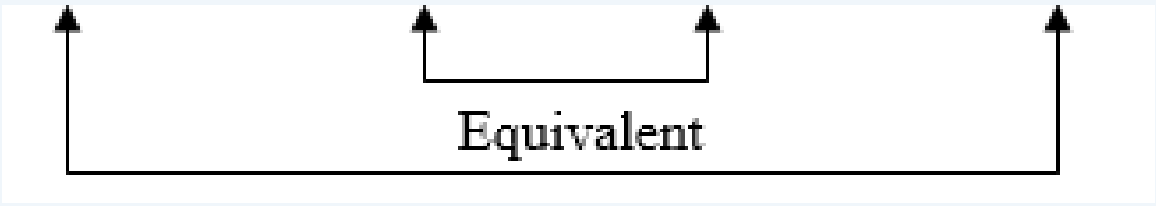
The inverse would be "If it is not raining, then there are not clouds in the sky." Likewise, this is not always true.

The contrapositive would be "If there are not clouds in the sky, then it is not raining." This statement is true, and is equivalent to the original conditional.

Looking at truth tables, we can see that the original conditional and the contrapositive are logically equivalent, and that the converse and inverse are logically equivalent.

Table 3.7.17

		Conditional	Converse	Inverse	Contrapositive
p	q	p → q	q → p	~p → ~q	~q → ~p
T	T	T	T	T	T
T	F	F	T	T	F
F	T	T	F	F	T
F	T	T	T	T	T



Equivalence

A conditional statement and its contrapositive are logically equivalent.

The converse and inverse of a conditional statement are logically equivalent.

In other words, the original statement and the contrapositive must agree with each other; they must both be true, or they must both be false. Similarly, the converse and the inverse must agree with each other; they must both be true, or they must both be false.

Be aware that symbolic logic cannot represent the English language perfectly. For example, we may need to change the verb tense to show that one thing occurred before another.

Example 3.7.10

Suppose this statement is true: "If I eat this giant cookie, then I will feel sick." Which of the following statements must also be true?

a. If I feel sick, then I ate that giant cookie.
b. If I don't eat this giant cookie, then I won't feel sick.
c. If I don't feel sick, then I didn't eat that giant cookie.

Solution

a. This is the converse, which is not necessarily true. I could feel sick for some other reason, such as drinking sour milk.

b. This is the inverse, which is not necessarily true. Again, I could feel sick for some other reason; avoiding the cookie doesn't guarantee that I won't feel sick.
c. This is the contrapositive, which is true, but we have to think somewhat backwards to explain it. If I ate the cookie, I would feel sick, but since I don't feel sick, I must not have eaten the cookie.

Notice again that the original statement and the contrapositive have the same truth value (both are true), and the converse and the inverse have the same truth value (both are false).

The Negation of a Conditional

The negation of a conditional statement is logically equivalent to a conjunction of the antecedent and the negation of the consequent.

~ (p → q) is equivalent to p ⧫ ~q

Example 3.7.11

Which of the following statements is equivalent to the negation of "If you don't grease the pan, then the food will stick to it"?

a. I didn't grease the pan and the food didn't stick to it.
b. I didn't grease the pan and the food stuck to it.
c. I greased the pan and the food didn't stick to it.

Solution

a. This is correct; it is the conjunction of the antecedent and the negation of the consequent. To disprove that not greasing the pan will cause the food to stick, I have to not grease the pan and have the food not stick.
b. This is essentially the original statement with no negation; the "if...then" has been replaced by "and".
c. This essentially agrees with the original statement and cannot disprove it.

In everyday life, we often have a stronger meaning in mind when we use a conditional statement. Consider "If you submit your hours today, then you will be paid next Friday." What the payroll rep really means is "If you submit your hours today, then you will be paid next Friday, and if you don't submit your hours today, then you won't be paid next Friday." The conditional statement if *t, then p* also includes the inverse of the statement: *if not t, then not p*. A more compact way to express this statement is "You will be paid next Friday *if and only if* you submit your timesheet today." A statement of this form is called a **biconditional.**

Biconditional

A biconditional is a logical conditional statement in which the antecedent and consequent are interchangeable.

A biconditional is written as p ↔ q and is translated as "p if and only if q".

Because a biconditional statement p ↔ q is equivalent to (p → q) ⧫ (q → p), we may think of it as a conditional statement combined with its converse: *if p, then q* and *if q, then p*. The double-headed arrow shows that the conditional statement goes from left to right and from right to left. A biconditional is considered true as long as the antecedent and the consequent have the same truth value; that is, they are either both true or both false.

Truth table for the biconditional

Table 3.7.18

p	q	p ↔ q
T	T	T
T	F	F
F	T	F
F	F	T

Notice that the fourth row, where both components are false, is true; if you don't submit your timesheet and you don't get paid, the person from payroll told you the truth.

Example 3.7.12

Suppose this statement is true: "The garbage truck comes down my street if and only if it is Thursday morning." Which of the following statements could be true?

a. It is noon on Thursday and the garbage truck did not come down my street this morning.
b. It is Monday and the garbage truck is coming down my street.
c. It is Wednesday at 11:59 PM and the garbage truck did not come down my street today.

Solution

a. This cannot be true. This is like the second row of the truth table; it is true that I just experienced Thursday morning, but it is false that the garbage truck came.
b. This cannot be true. This is like the third row of the truth table; it is false that it is Thursday, but it is true that the garbage truck came.
c. This could be true. This is like the fourth row of the truth table; it is false that it is Thursday, but it is also false that the garbage truck came, so everything worked out like it should.

Example 3.7.13

Create a truth table for the statement (A ⋄ B) ↔ ~C

Solution

Whenever we have three component statements, we start by listing all the possible truth value combinations for A, B, and C. After creating those three columns, we can create a fourth column for the antecedent, A ⋄ B. Now we will temporarily ignore the column for C and focus on A and B, writing the truth values for A ⋄ B.

Table 3.7.19

A	B	C
T	T	T
T	T	F

A	B	C
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

Table 3.7.20

A	B	C	A ⊘ B
T	T	T	T
T	T	F	T
T	F	T	T
T	F	F	T
F	T	T	T
F	T	F	T
F	F	T	F
F	F	F	F

Next we can create a column for the negation of C. (Ignore the A ⧫ B column and simply negate the values in the C column.)

Table 3.7.21

A	B	C	A ⧫ B	~C
T	T	T	T	F
T	T	F	T	T
T	F	T	T	F
T	F	F	T	T
F	T	T	T	F
F	T	F	T	T
F	F	T	F	F
F	F	F	F	T

Finally, we find the truth values of (A ⧫ B) ↔ ~C. Remember, a biconditional is true when the truth value of the two parts match, but it is false when the truth values do not match.

Table 3.7.22

A	B	C	A ⧫ B	~C	(A ⧫ B) ↔ ~C
T	T	T	T	F	F
T	T	F	T	T	T
T	F	T	T	F	F
T	F	F	T	T	T

A	B	C	A ⊽ B	~C	(A ⊽ B) ↔ ~C
F	T	T	T	F	F
F	T	F	T	T	T
F	F	T	F	F	T
F	F	F	F	T	F

To illustrate this situation, suppose your boss needs you to do either project A or project B (or both, if you have the time). If you do one of the projects, you will not get a crummy review (C is for crummy). So (A ⊽ B) ↔ ~C means "You will not get a crummy review if and only if you do project A or project B." Looking at a few of the rows of the truth table, we can see how this works out. In the first row, A, B, and C are all true: you did both projects and got a crummy review, which is not what your boss told you would happen! That is why the final result of the first row is false. In the fourth row, A is true, B is false, and C is false: you did project A and did not get a crummy review. This is what your boss said would happen, so the final result of this row is true. And in the eighth row, A, B, and C are all false: you didn't do either project and did not get a crummy review. This is not what your boss said would happen, so the final result of this row is false. (Even though you may be happy that your boss didn't follow through on the threat, the truth table shows that your boss lied about what would happen.)

De Morgan's Laws

A contemporary of Boole's, Augustus De Morgan, formalized two rules of logic that had previously been known informally. They allow us to rewrite the negation of a conjunction as a disjunction, and vice-versa.

For example, suppose you want to schedule a meeting with two colleagues at 4:30 PM on Friday, and you need both of them to be available at that time. What situation would make it impossible to have the meeting? It is NOT the case that colleague a is available AND colleague *b* is available: ~ (a ∧ b). This situation is equivalent to either colleague a NOT being available OR colleague b NOT being available: ~a ∨ ~b.

De Morgan's Laws

The negation of a conjunction is equivalent to the disjunction of the negation of the statements making up the conjunction. To negate an "and" statement, negate each part and change the "and" to "or".

$$\sim (p \wedge q) \text{ is equivalent to } \sim p \wedge \sim q$$

The negation of a disjunction is equivalent to the conjunction of the negation of the statements making up the disjunction. To negate an "or" statement, negate each part and change the "or" to "and".

$$\sim (p \vee q) \text{ is equivalent to } \sim p \vee \sim q$$

Example 3.7.14

For Valentine's Day, you did not get your sweetie flowers or candy. Which of the following statements is logically equivalent?

a. You did not get them flowers or did not get them candy.
b. You did not get them flowers and did not get them candy.
c. You got them flowers or got them candy.

Solution

a. This statement does not go far enough; it leaves open the possibility that you got them one of the two things.
b. This statement is equivalent to the original; ~(f ∨ c) is equivalent to ~f ∧~c.
c. This statement says that you got them something, but we know that you did not.

Contributors and Attributions

- Saburo Matsumoto
  CC-BY-4.0

This page titled 3.7: Truth Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by LibreTexts.

3.8: Analyzing Arguments with Truth Tables

Learning Objectives

Students will be able to:

- Determine the validity of an argument using truth tables

The validity of arguments also be determined using truth tables.

Analyzing arguments using truth tables

To analyze an argument with a truth table:

1. Represent each of the premises symbolically.
2. Create a conditional statement, joining all the premises to form the antecedent and using the conclusion as the consequent.
3. Create a truth table for the statement. If it is always true, then the argument is valid.

Example 3.8.6

Consider the argument

Premise: If you bought bread, then you went to the store.

Premise: You bought bread.

Conclusion: You went to the store.

Solution

While this example is fairly obviously a valid argument, we can analyze it using a truth table by representing each of the premises symbolically. We can then form a conditional statement showing that the premises together imply the conclusion. If the truth table is a tautology (always true), then the argument is valid.

We'll let b represent "you bought bread" and s represent "you went to the store". Then the argument becomes:

Premise: $b \rightarrow s$

Premise: b

Conclusion: s

To test the validity, we look at whether the combination of both premises implies the conclusion; is it true that [(b → s) ◈ b] → s ?

Table 3.8.1

b	s	b → s
T	T	T
T	F	F
F	T	T
F	F	T

Table 3.8.2

b	s	b → s	(b → s) ⧑ b
T	T	T	T
T	F	F	F
F	T	T	F
F	F	T	F

Table 3.8.3

b	s	b → s	(b → s) ⧑ b	[(b → s) ⧑ b] → s
T	T	T	T	T
T	F	F	F	T
F	T	T	F	T
F	F	T	F	T

Since the truth table for [(b → s) ⋄ b] → s is always true, this is a valid argument.

Example 3.8.7

Premise: If I go to the mall, then I'll buy new jeans.

Premise: If I buy new jeans, I'll buy a shirt to go with it.

Conclusion: If I go to the mall, I'll buy a shirt.

Solution

Let *m* = I go to the mall, *j* = I buy jeans, and s = I buy a shirt.

The premises and conclusion can be stated as:

Premise: $m \rightarrow j$

Premise: $j \rightarrow s$

Conclusion: $m \rightarrow s$

We can construct a truth table for [(m → j) ⋄ (j → s)] → (m → s). Try to recreate each step and see how the truth table was constructed.

Table 3.8.4

m	j	s	m→j	j→s	(m→j) ⋄ (j→s)	m→s	[(m→j) ⋄ (j→s)] → (m→s)
T	T	T	T	T	T	T	T
T	T	F	T	F	F	F	T
T	F	T	F	T	F	T	T
T	F	F	F	T	F	F	T
F	T	T	T	T	T	T	T
F	T	F	T	F	F	T	T
F	F	T	T	T	T	T	T
F	F	F	T	T	T	T	T

From the final column of the truth table, we can see this is a valid argument.

Contributors and Attributions

- Saburo Matsumoto
  CC-BY-4.0

This page titled 3.8: Analyzing Arguments with Truth Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by LibreTexts.

4: DESCRIPTIVE STATISTICS



- Topic hierarchy

4.1: Introduction to Statistics

4.2: Collecting Data

4.3: Visual Summaries of Data

4.4: Other Summaries

4.5: Measures of the Center and Variation of the Data

4.6: Measures of the Location of the Data and Outliers

This page titled 4: Descriptive Statistics is shared under a license and was authored, remixed, and/or curated by LibreTexts.

4.1: Introduction to Statistics

Learning Objectives

- Recognize and differentiate between key terms.
- Identify different types of data

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

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

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

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

Key Terms in Statistics

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

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

A **parameter** is a number that represents a property of the *population*. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter. A **statistic** is a number that represents a property of the *sample*. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. A statistic can be used to **estimate** the unknown population parameters.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's political party affiliation, then some examples of Y include Republican, Democrat, and Independent; Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y

(calculating an average party affiliation makes no sense).

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

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 rounded to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Example 4.1.1

Determine what the key terms (population, sample, parameter, statistics, variable, data) refer to in the following study. We want to know the average (mean) amount of money first-year college students spend at MSJC on school supplies that do not include books. We randomly survey 100 first-year students at the college. Three of those students spent $150, $200, and $225, respectively.

Answer

- The **population** is all first-year students attending AMSJC this term.
- The **sample** could be all students enrolled in one section of a beginning statistics course at MSJC (although this sample may not represent the entire population).
- The **parameter** is the average (mean) amount of money spent (excluding books) by first-year college students at MSJC this term.
- The **statistic** is the average (mean) amount of money spent (excluding books) by first-year college students in the sample.
- The **variable** could be the amount of money spent (excluding books) by one first-year student. Let X = the amount of money spent (excluding books) by one first-year student attending MSJC.
- The **data** are the dollar amounts spent by the first-year students. Examples of the data are $150, $200, and $225.

Exercise 4.1.1

Determine what the key terms (population, sample, parameter, statistics, variable, data) refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95, respectively.

Answer available at the end of this section.

Example 4.1.2

Determine what the key terms (population, sample, parameter, statistics, variable, data) refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

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

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries if they had been actual drivers. We start with a simple random sample of 75 cars.

Answer

- The **population** is all cars containing dummies in the front seat.
- The **sample** is the 75 cars, selected by a simple random sample.
- The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.
- The **statistic** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.
- The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.
- The **data** are either: yes, had a head injury, or no, did not.

Exercise 4.1.2

Determine what the key terms (population, sample, parameter, statistics, variable, data) refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Answer available at the end of this section.

Types of Data

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

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of *counting* or *measuring* attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either *discrete* or *continuous*. *Quantitative discrete data* takes on only certain numerical values or in other words it would be possible to list all possible values. For example, if you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three, thus the number of daily calls is discrete data. All data that are the result of measuring are *quantitative continuous data* assuming that we can measure accurately. Measuring lengths in yards might result in such numbers as 10.2, 7.9, 102.6, 92.7, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Example 4.1.1

Consider the number of books students carry in their backpacks in a sample of five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. What type of data is this?

Solution

The numbers of books (three, four, two, and one) are **quantitative discrete data**.

Exercise 4.1.1

Consider the number of machines in a gym in a sample of five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Answer available at the end of this section.

Example 4.1.2

You go to the supermarket and purchase three cans of soup (19.4 ounces) tomato bisque, 14.1 ounces lentil, and 19.6 ounces Italian wedding, two packages of nuts (walnuts and peanuts), four different kinds of vegetables (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and two pounds (32 ounces chocolate chip cookies). What type of data are the weights of the soups?

Solution

The weights of the soups (19.4 ounces, 14.1 ounces, 19.6 ounces) are quantitative continuous data because you measure weights as precisely as possible.

Exercise 4.1.2

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Answer available at the end of this section.

Example 4.1.3

Consider the colors of backpacks of a sample of five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. What type of data is this?

Solution

The colors red, black, black, green, and gray are **qualitative data**.

Exercise 4.1.3

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Answer available at the end of this section.

Interactive Exercise 4.1.4

Answers

Exercise 4.1.1

- The **population** is all families with children attending Knoll Academy.
- The **sample** is a random selection of 100 families with children attending Knoll Academy.
- The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.
- The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.
- The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll

Academy.

- The **data** are the dollar amounts spent by the families. Examples of the data are $65, $75, and $95.

Exercise 4.1.2

- The **population** is all medical doctors listed in the professional directory.
- The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.
- The **sample** is the 500 doctors selected at random from the professional directory.
- The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.
- The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.
- The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Exercise 4.1.1

The numbers of machines (three, four, two, and one) are quantitative discrete data.

Exercise 4.1.2

quantitative continuous data

Exercise 4.1.3

qualitative data

This page titled 4.1: Introduction to Statistics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

4.2: Collecting Data

Learning Objectives

- Identify various types of sampling methods of data collection.

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. Other well-known random sampling methods are **the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However, for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only

once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low. However, if the sample size is less than 5% of the population then the difference between sampling with replacement or without is insignificant.

When you analyze data, it is important to be aware of **sampling errors** and **non-sampling** errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause non-sampling errors. A defective counting device can cause a non-sampling error. In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example 4.2.1

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

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high-tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on average.

Answer

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f. convenience

Exercise 4.2.1

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Answer

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example 4.2.2

Suppose MSJC has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task. Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first-term organic chemistry class. Many of these students are taking first-term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128; $87; $173; $116; $130; $204; $147; $189; $93; $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50; $40; $36; $15; $50; $100; $40; $53; $22; $22

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

c. Is the sample biased?

Answer

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180; $50; $150; $85; $260; $75; $180; $200; $200; $150

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population. Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Exercise 4.2.2

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Answer

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

Interactive Exercise 4.2.2

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

View this Interactive Content online



Variation in Samples and Size of a Sample

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16-ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however. If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This *variability in samples* cannot be stressed enough.

The **size of a sample** (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals. Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Critical Evaluation

We need to critically evaluate the statistical studies we read about and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions.
- Undue influence: collecting data or asking questions in a way that influences the response can make the data unreliable.
- Non-response or refusal of the subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context.

- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Ethics

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

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

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

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot. Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[3] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis. Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data. When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. These ideas may seem fundamental, but they can be very difficult to verify in practice.

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

Example 4.2.3

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

A researcher is collecting data in a community.

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

Answer

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

Exercise 4.2.3

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

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

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

Answer

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

Contributors and Attributions

- Template:ContribOpenStax

This page titled 4.2: Collecting Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

4.3: Visual Summaries of Data

Learning Objectives

- Create and interpret frequency tables.
- Create and interpret bar plots and histograms.

Organizing Data

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

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

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

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

Frequency Tables

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table lists the different data values in ascending order and their frequencies.

Table $4.3.A$*: Frequency Table of Student Work Hours*

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

Definition: Relative Frequency

A **frequency** is the number of times a value of the data occurs. According to Table Table $4.3.A$, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

Definition: Relative frequencies

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

Table $4.3.B$: Frequency Table of Student Work Hours with Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15 or 15%
3	5	$\frac{5}{20}$ or 0.25 or 25%
4	3	$\frac{3}{20}$ or 0.15 or 15%
5	6	$\frac{6}{20}$ or 0.30 or 30%
6	2	$\frac{2}{20}$ or 0.10 or 10%
7	1	$\frac{1}{20}$ or 0.05 or 5%

The sum of the values in the relative frequency column of Table $4.3.B$ is $\frac{20}{20}$, or 1, or 100%.

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they should be close to one.

Table $4.3.C$ represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table $4.3.C$: Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY
60–62	5	$\frac{5}{100} = 0.05 = 5$
62-64	3	$\frac{3}{100} = 0.03 = 3$
64–66	15	$\frac{15}{100} = 0.15 = 15$
66–68	40	$\frac{40}{100} = 0.40 = 40$
68–70	17	$\frac{17}{100} = 0.17 = 17$
70–72	12	$\frac{12}{100} = 0.12 = 12$
72–74	7	$\frac{7}{100} = 0.07 = 7$
74–76	1	$\frac{1}{100} = 0.01 = 1$
	Total = 100	**Total = 1.00**

In this sample, there are **five** players whose heights fall within the interval 60–62 inches, **three** players whose heights fall within the interval 62–64 inches, **15** players whose heights fall within the interval 64–66 inches, **40** players whose heights fall within the interval 66–68 inches, **17** players whose heights fall within the interval 68–70 inches, **12** players whose heights fall within the interval 70–72, **seven** players whose heights fall within the interval 72–74, and **one** player whose heights fall within the interval 74–76. All heights fall between the endpoints of an interval and not at the endpoints.

Example 4.3.1

a. From the Table $4.3.C$, find the percentage of heights that are less than 66 inches.
b. Find the percentage of heights that fall between 62 and 66 inches.

Answer

a. If you look at the first, second, and third rows, the heights are all less than 66 inches. There are $5 + 3 + 15 = 23$ players whose heights are less than 66 inches. The percentage of heights less than 66 inches is then $\frac{23}{100}$ or 23%. This percentage is an example of the **cumulative relative frequency**.
b. Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$ or 18%.

Exercise 4.3.1

Table $4.3.D$ shows the amount, in inches, of annual rainfall in a sample of towns.

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

Table $4.3.D$

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

Answer available at the end of this section.

Exercise 4.3.2

Use the heights of the 100 male semiprofessional soccer players in Table $4.3.C$. Fill in the blanks and check your answers.

a. The percentage of heights that are from 68 to 72 inches is: ____.
b. The percentage of heights that are from 68 to 74 inches is: ____.
c. The percentage of heights that are more than 66 inches is: ____.
d. The number of players in the sample who are between 62 and 72 inches tall is: ____.
e. What kind of data are the heights?
f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Answer

a. 29%

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

Exercise 4.3.3

From Table 4.3.D, find the number of towns that have rainfall between 2.95 and 9.01 inches.

Answer available at the end of this section.

Example 4.3.4

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Table 4.3.E was produced:

Table 4.3.E: Frequency of Commuting Distances

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

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

Answer

a. No. The frequency column sums to 18, not 19.
b. False. The frequency for three miles should be one; for two miles (left out), two.
c. $\frac{5}{19}$
d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Exercise 4.3.4

Table 4.3.D represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Answer available at the end of this section.

Example 4.3.5

Table 4.3.F contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Table 4.3.F: Total Number of Deaths Worldwide as a Result of Earthquakes

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,356

Answer the following questions.

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

Answer

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

Exercise 4.3.5

Table 4.3.G contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Table 4.3.G:

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the percentage of deaths that occurred between 1994 and 2006? Explain what this number tells you about the data.

Answer

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

Visualizing Data

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

Table 4.3.H: Fall Term 2007 (Census day)

De Anza College				Foothill College		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%

De Anza College				Foothill College		
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

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

Look at Figures $4.3.I$ and $4.3.J$ and determine which graph (pie or bar) you think displays the comparisons better.

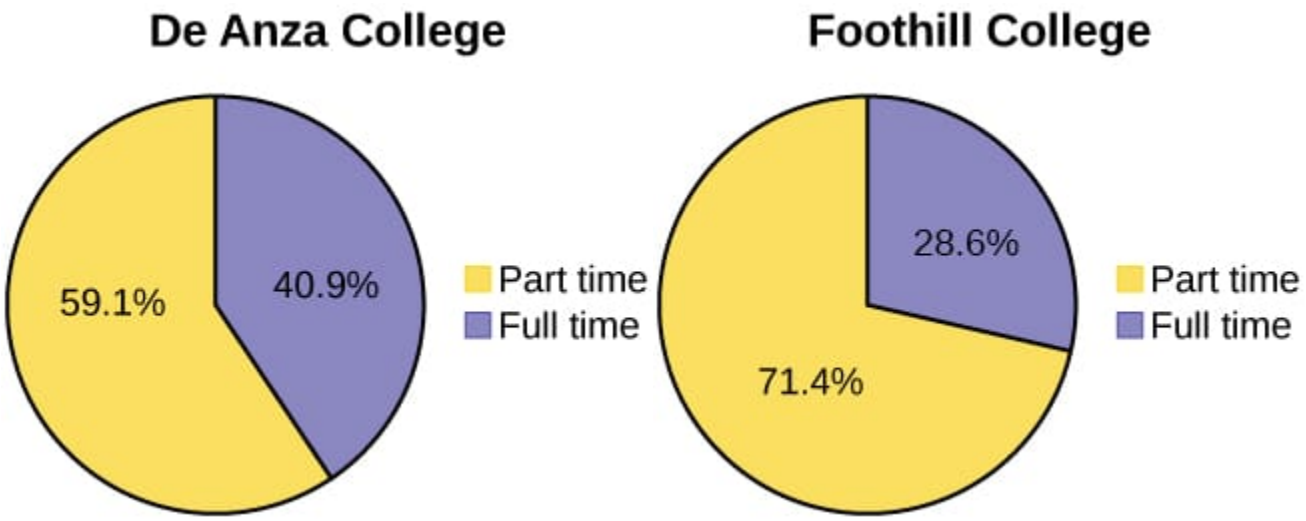


Figure $4.3.I$: Pie Charts

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

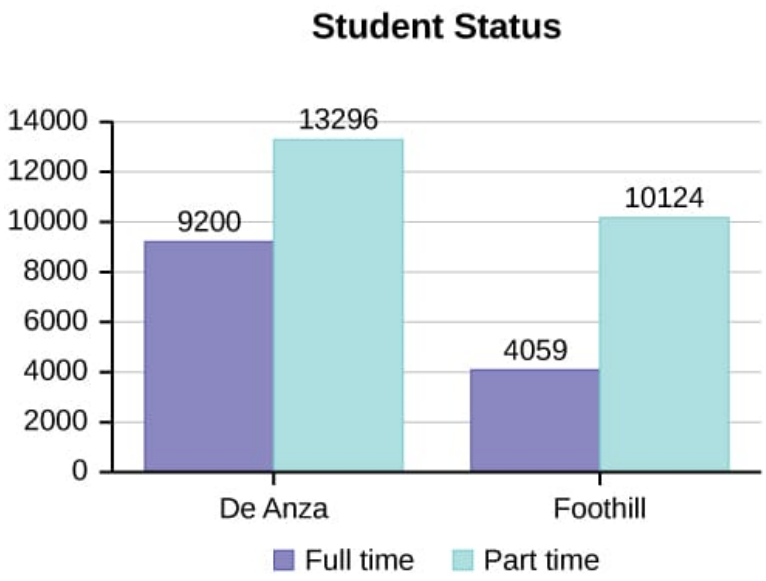


Figure 4.3.J: Bar chart

Bar Graphs

Bar graphs consist of bars that are separated from each other. The bars can be rectangles, or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graphs** have the possible values represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 4.3.6

By the end of 2011, Facebook had over 146 million users in the United States. Table shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Answer

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

Figure 4.3.K.

Exercise 4.3.6

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

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

Age groups	Number of people	Proportion of population
Retirees	131,662	38%

Answer

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

Figure 4.3.*L*.

Example 4.3.7

The columns in the following table contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Solution

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

Figure 4.3.*M*

Exercise 4.3.7

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Answer

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

Figure $4.3.N$

Histograms

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

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

Example $4.3.8$

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

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

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

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

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

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

72; 72; 72; 72.5; 72.5; 73; 73.5

74

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

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

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

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

$$\frac{74.05 - 59.95}{8} = 1.76 \tag{4.3.1}$$

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

The boundaries are:

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

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

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

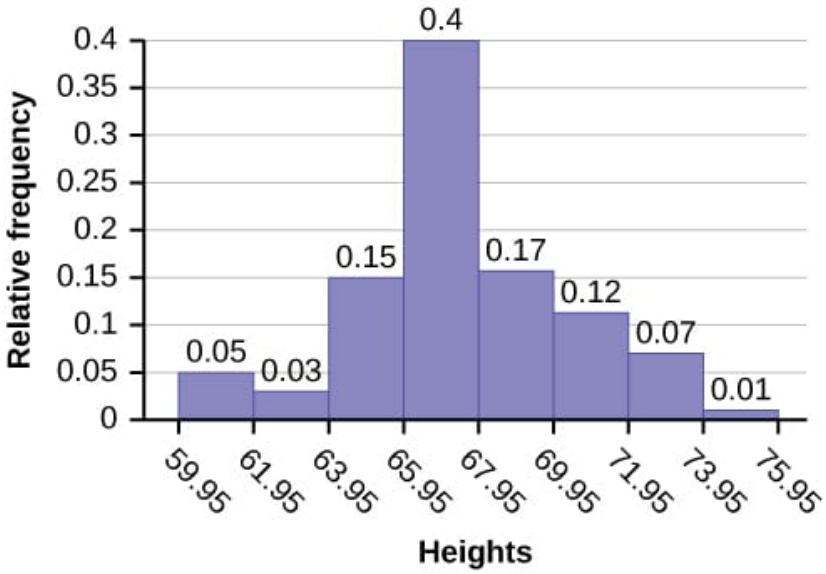


Figure 4.3.*O*

Exercise 4.3.8

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

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

Answer

Smallest value: 9

Largest value: 14

Convenient starting value: 9 – 0.05 = 8.95

Convenient ending value: 14 + 0.05 = 14.05

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

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

Example 4.3.9

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

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

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

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

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

Answer

Calculate the number of bars as follows:

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

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

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

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

Figure 4.3.*P*.

Exercise 4.3.9

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

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

2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3

20 student athletes play one sport. 22 student athletes play two sports. Eight student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

Answer

1.5

1.5 to 2.5

2.5 to 3.5

Example 4.3.10

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends

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

Answer

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

Figure 4.3.Q.

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

Exercise 4.3.10

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

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

Use 10–19 as the first interval.

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

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A

pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Table 4.3.R: De Anza College Spring 2010

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

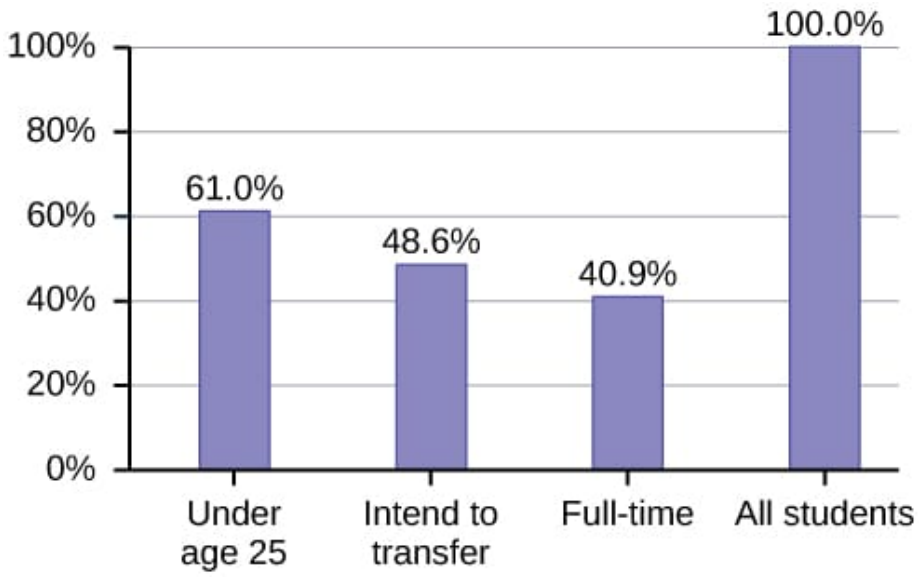


Figure 4.3.S: Bar chart of data in Table 4.3.R.

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

Table 4.3.T: Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

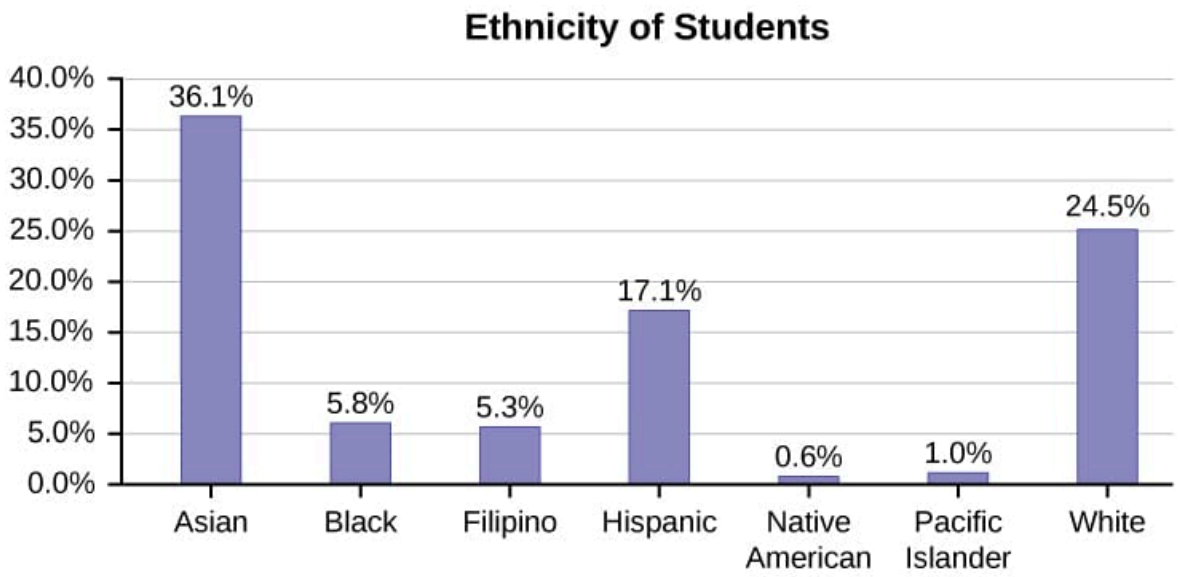


Figure 4.3.U: Enrollment of De Anza College (Spring 2010)

The following graph is the same as the previous graph, but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure 4.3.4 can be difficult to understand visually. The graph in Figure 4.3.5 is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

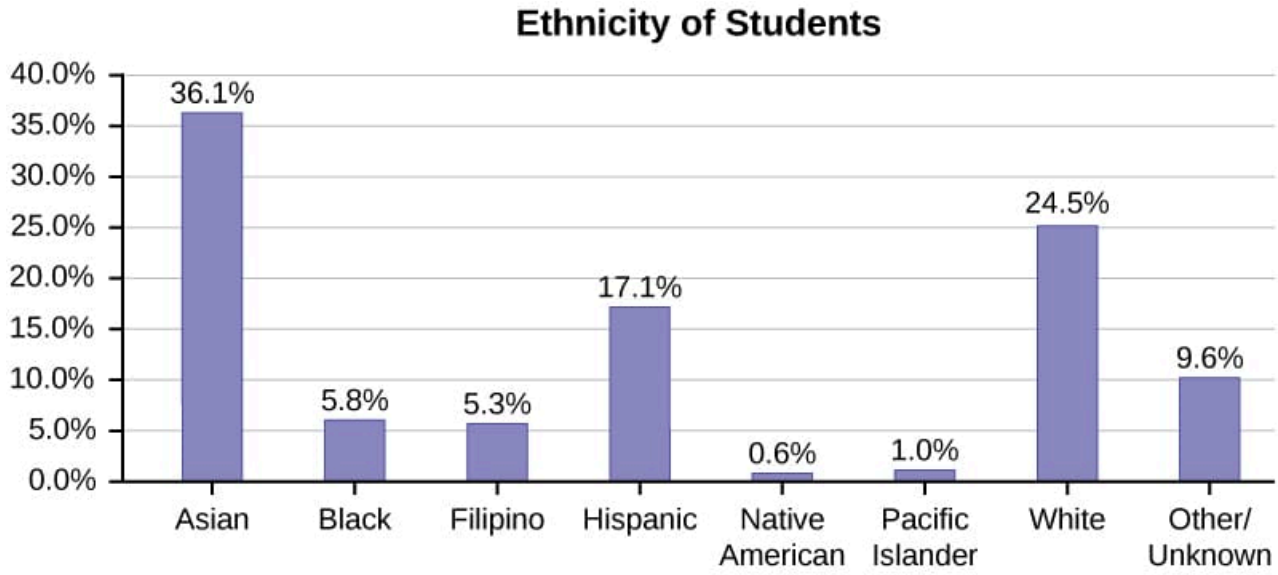


Figure 4.3.V: Bar Graph with Other/Unknown Category

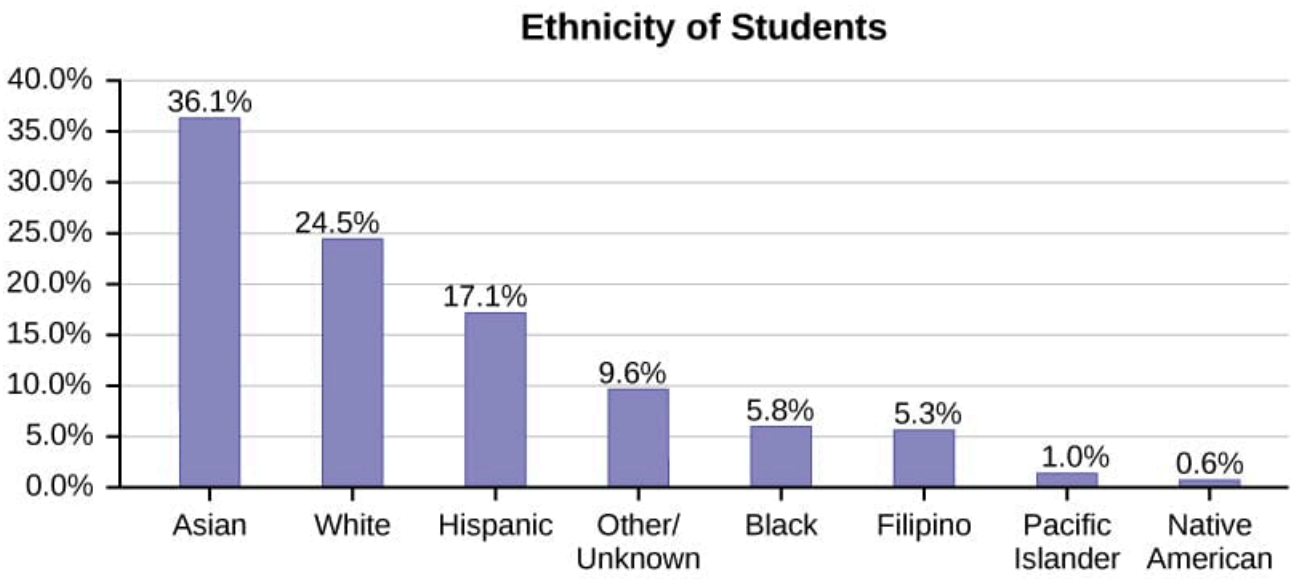


Figure $4.3.W$: Pareto Chart With Bars Sorted by Size

Contributors and Attributions

- Template:ContribOpenStax

Answers

Exercise $4.3.1$

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

Exercise $4.3.3$

$6 + 7 + 15 = 28$ towns

Exercise $4.3.4$

$\frac{9}{50}$

This page titled 4.3: Visual Summaries of Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

4.4: Other Summaries

Learning Objectives

- Create and interpret stemplots, line graphs, time series for univariate data.
- Create and interpret contingency tables and scatter plots for bivariate data.

Stem-and-Leaf Diagrams or Stemplots

One simple graph, the *stem-and-leaf graph* or *stemplot*, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a final significant digit. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 4.4.1

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

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

Stem-and-Leaf Graph

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

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Exercise 4.4.1

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

Answer available at the end of this section.

Interactive Exercise 4.4.1

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

View this Interactive Content online

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Example 4.4.2

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

HINT: The leaves are to the right of the decimal.

Answer

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Exercise 4.4.2

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Answer available at the end of this section.

Example 4.4.3: Side-by-Side Stem-and-Leaf plot

A side-by-side stem-and-leaf plot allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Tables 4.4.1 and 4.4.2 show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Table 4.4.1: Presidential Ages at Inauguration

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

4.4.2 *Presidential Age at Death*

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81

President	Age	President	Age	President	Age
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Answer

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

Exercise 4.4.3

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

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003

Losses	Wins	Year	Losses	Wins	Year
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Answer available at the end of this section.

Line Graphs or Frequency Polygons

Another type of graph that is useful for specific data values is a **line graph** or a **frequency polygon**. In the particular line graph shown in the example below, the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

Example 4.4.4

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table and in Figure.

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

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

Figure 4.4.1

Exercise 4.4.4

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table. Construct a line graph.

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

Answer

Figure 4.4.2.

Interactive Exercise 4.4.4

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

View this Interactive Content online

Frequency polygons can also be used for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example 4.4.5

We will construct an overlay frequency polygon comparing the final exam scores with the students' final numeric grade.

Frequency Distribution for Calculus Final Test Scores

Range	Midpoint	Frequency	Cumulative Frequency
49.5-59.5	54.5	5	5
59.5-69.5	64.5	10	15
69.5-79.5	74.5	30	45
79.5-89.5	84.5	40	85
89.5-99.5	94.5	15	100

Frequency Distribution for Calculus Final Grades

Range	Midpoint	Frequency	Cumulative Frequency
49.5-59.5	54.5	10	10
59.5-69.5	64.5	10	20
69.5-79.5	74.5	30	50
79.5-89.5	84.5	45	95
89.5-99.5	94.5	5	100

This is an overlay frequency polygon that matches the supplied data. The x-axis shows the grades, and the y-axis shows the frequency.

Figure 4.4.6.

Time Series Graphs

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

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

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

Example 4.4.6

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

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

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Answer

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

Figure 4.4.7.

Exercise 4.4.6

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

CO_2 Emissions

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

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

Figure 4.4.8.

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

Bivariate Data

Earlier we discussed the methods of summarizing and organizing the data obtained from observing one variable. Such data are called **univariate data**. When we are interested in studying the relationship between a pair of variables we must collect and organize the data on two variables at the same time. Such data itself is called **bivariate data** and there are several ways to organize such data depending on the types of the variables.

Univariate Data	Bivariate Data

The mileage of 5 cars

Car	Mileage (mi)
1	78524
2	12574
3	24914
4	65813
5	39824

The ages and mileage of 5 cars

Car	Age (years)	Mileage (mi)
1	7	78524
2	2	12574
3	1	24914
4	5	65813
5	3	39824

Scatter Plots

When both variables are quantitative, we can construct a scatter plot to organize the data. To construct a **scatterplot**, we use a horizontal axis for the observations of one variable and a vertical axis for the observations of the other. When picking which axis to use for each variable consider whether you suspect that one variable depends on the other. The independent variable will be on the x-axis and dependent will be on the y axis. Each pair of observations is then plotted as a point.

Example 4.4.7

The summary of the ages and the mileages of a sample of 5 cars is shown below:

Table 4.4.A: Table showing the ages and the mileages of a sample of 5 cars.

Car	x (age in years)	y (mileage in miles)
1	7	78524
2	2	12574
3	1	24914
4	5	65813

Car	x (age in years)	y (mileage in miles)
5	3	39824

To construct the **scatterplot**, we use a horizontal axis for the ages of cars and a vertical axis for the mileage.

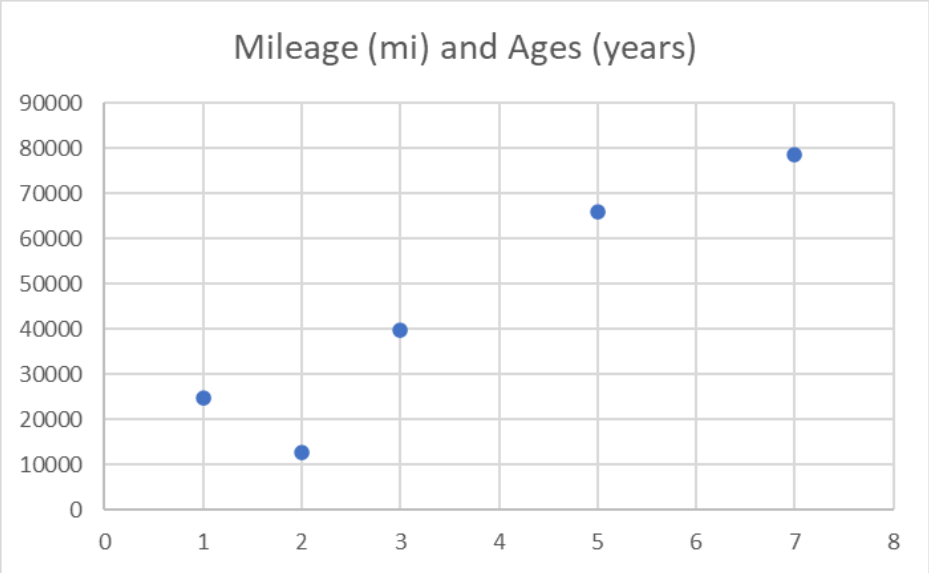


Figure $4.4.B$: Scatter plot showing the ages and the mileages of a sample of 5 cars.

Exercise 4.4.7

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Answer available at the end of this section.

Interactive Exercise 4.4.7

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

View this Interactive Content online

Contingency Tables

When we are interested in studying the relationship between a pair of variables we must collect and organize the data on two variables at the same time. Recall that such data itself is called **bivariate data** and there are several ways to organize such data depending on the types of the variables. Below are examples of univariate and bivariate qualitative data.

Univariate Data

The colors of 5 cars

Car	Color
1	Light
2	Dark
3	Dark
4	Light
5	Dark

Bivariate Data

The color and condition of 5 cars

Car	Color	Condition
1	Light	Like new
2	Dark	Used
3	Dark	Like New
4	Light	Used
5	Dark	Used

A table called **contingency table** can be used to organize bivariate data in which one or both variables are qualitative.

Example 4.4.8

To construct the table, we arrange the observed frequencies into rows and columns. The intersection of a row and a column of a contingency table is called a cell. Each cell shows the frequency of observations that fit the description in the corresponding row and column.

Color\Condition	Like new	Used
Light	1	1
Dark	1	2

Many times, it makes sense to add the column and row with the totals for each column and row and for the entire table.

Color\Condition	Like new	Used	Total:
Light	1	1	2
Dark	1	2	3
Total:	2	3	5

Alternatively, the contingency table can be referred to as a **two-way frequency table**. Why? In the example above, cover the second and third columns and what's left can be easily seen as the frequency table for the colors of cars! Cover the second and third rows instead and what's left can be easily seen as the frequency table for whether car conditions! Also note that every contingency table can be easily broken down into two frequency tables, but the two frequency tables cannot be combined into a contingency table if the original data is lost!

Exercise 4.4.8

Destiny surveyed the students in her neighborhood and obtained the following data. Construct the contingency table that summarizes the school enrollment by level and type.

Student	School level	School type
Student 1	Middle School	Public
Student 2	Middle School	Private
Student 3	Elementary School	Public
Student 4	Middle School	Public
Student 5	Elementary School	Public
Student 6	High School	Public
Student 7	High School	Private
Student 8	Middle School	Public
Student 9	High School	Public
Student 10	Elementary School	Public

Student	School level	School type
Student 11	Middle School	Private
Student 12	Elementary School	Private
Student 13	Elementary School	Private
Student 14	Elementary School	Private
Student 15	Elementary School	Private
Student 16	Middle School	Private
Student 17	Middle School	Private
Student 18	High School	Private
Student 19	Elementary School	Private

Answer available at the end of this section.

Exercise 4.4.8

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

View this Interactive Content online

Answers

Exercise 4.4.1

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

Exercise 4.4.2

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

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

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

Exercise 4.4.3

Table 4.4.A: Atlanta Hawks Wins and Losses

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

Exercise 4.4.7

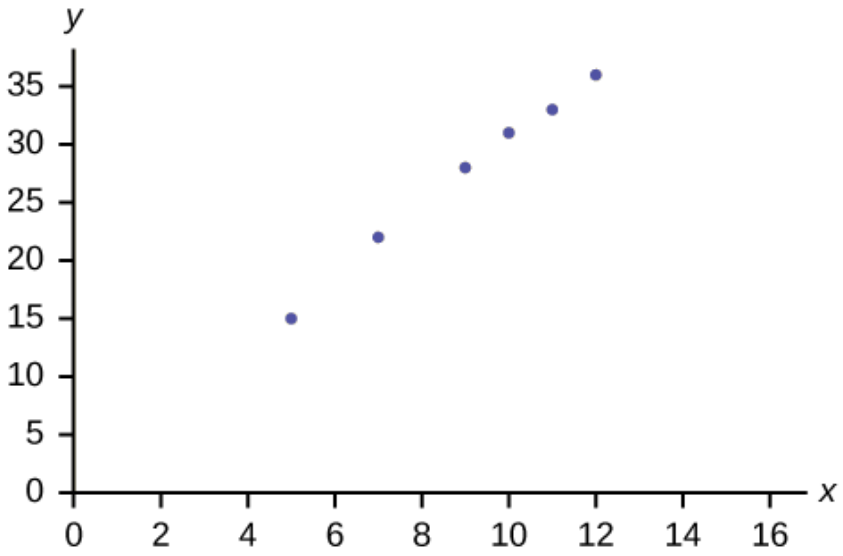


Figure 4.4.2

Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

Exercise 4.4.8

The following contingency table summarizes the school enrollment by level and type:

	Public	Private	Total
Elementary School	3	5	8
Middle School	3	4	7
High School	2	2	4
Total	8	11	19

This page titled 4.4: Other Summaries is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.
- 11.1: Self-adjoint or hermitian operators by Isaiah Lankham, Bruno Nachtergaele, and Anne Schilling has no license indicated.

4.5: Measures of the Center and Variation of the Data

Learning Objectives

- Recognize, describe, and calculate the measures of the center of data.
- Recognize, describe, and calculate the measures of the spread of data.
- Use Chebyshev and Empirical rules to interpret the mean and standard deviation.

Measures of the Center of the Data

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

The letter used to represent the sample mean is an x with a bar over it (pronounced "x bar"): $\overline{x}$. The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

The Law of Large Numbers and the Mean

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

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. To see that both ways of calculating the mean are the same, consider the sample:

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

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{4.5.1}$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7 \tag{4.5.2}$$

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

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

$$\frac{n+1}{2} \tag{4.5.3}$$

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

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \tag{4.5.4}$$

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

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \tag{4.5.5}$$

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

Example 4.5.1

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

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

Calculate the mean and the median.

Answer

The calculation for the mean is:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6 \tag{4.5.6}$$

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

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \tag{4.5.7}$$

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

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

$$M = \frac{24+24}{2} = 24 \tag{4.5.8}$$

Exercise 4.5.1

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

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

Answer

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

$$\frac{544}{39} = 13.95 \tag{4.5.9}$$

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

Interactive Exercise 4.5.1

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

View this Interactive Content online

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

Example 4.5.2

Statistics exam scores for 20 students are as follows:

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

Find the mode.

Answer

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

Exercise 4.5.2

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

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

Find the mode.

Answer

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

Interactive Exercise 4.5.2

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

View this Interactive Content online

It is worth mentioning that the mean and median apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative data. Thus while the mode appears to be the least useful among the three measures of central tendency, it is the only way to assess the center for qualitative data! But when working with quantitative data we still have to choose between the mean and the median. Let's consider the following example. We have four people Ann, Beth, Connor, and Diana that make the annual salaries $30k, $50k, $50k, $70k respectfully. Their average salary is $50,000 and their median is also $50,000 dollars. Now, let's say Elon moves in and his salary is $300,000 a year. What effect does this have on the mean and median? The new mean is $100,000 and the new median is still $50,000. Which of the two measures produces a more representative value? In this case, the median is obviously a better choice, for example if Toyota Company was considering building a Lexus or Scion dealership and they were only considering the mean it would appear that an average person in this neighborhood can afford a Lexus when in reality only one person can afford it. So what exactly the difference between the mean and the median? Unlike the mean, the median is not sensitive to the influence of a few extreme observations. Usually, we use the median for house prices and salaries and for everything else we would prefer the mean.

Example 4.5.3

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

Solution

$$\bar{x} = \frac{5,000,000 + 49(30,000)}{50} = 129,400 \tag{4.5.10}$$

$M = 30,000$

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

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

Exercise 4.5.3

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

Answer

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

Interactive Exercise 4.5.3

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

View this Interactive Content online

Measures of Variation of the Data

One of the differences between the two data sets that any measure of center doesn't capture is the variety of data within the set. To describe the variation quantitatively, we use **measures of variation** or **measures of spread**. Just as there are several different measures of center, there are also several different measures of variation. In this section, we examine two of the most frequently used measures of variation: the range and standard deviation.

Definition: Range

The **range** of a data set is the difference between the maximum (largest) and minimum (smallest) observations.

Example 4.5.4

Find the range of the data:

8	12	13	11	10	9	14	8	6	14	7	8	13

Solution

The range of the data is the difference between the largest and the smallest values in the data set: 14−6=8

Interactive Exercise 4.5.4

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

View this Interactive Content online

Definition: The Standard Deviation

The range only measures the total variation and doesn't capture any variation between the minimum and maximum observed values. In contrast to the range, the **standard deviation** takes into account all the observations. It is the preferred measure of variation when the mean is used as the measure of center. Roughly speaking, the standard deviation measures variation by indicating how far, on average, the observations are from the mean. For a data set with a large amount of variation, the observations will, on average, be far from the mean; so the standard deviation will be large. For a data set with a small amount of variation, the observations will, on average, be close to the mean; so the standard deviation will be small.

Calculating the Standard Deviations

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

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

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

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

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{4.5.11}$$

or

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \tag{4.5.12}$$

For the sample standard deviation, the denominator is $n - 1$, that is one less than the sample size.

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{4.5.13}$$

or

$$\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}} \tag{4.5.14}$$

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

In Equations 4.5.12 and 4.5.14, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Why divide by n-1?

Why divide by (n-1) instead of n? Remember that the main purpose of finding a statistic is to estimate the unknown parameter. If we divide by n instead of (n-1) we will be obtaining a value which is a biased estimator for the population variance. Being biased doesn't necessarily mean bad when we know what the bias is. In this case, \sqrt{\dfrac{\sum(x-\mu)^{2}}{n}} always underestimates the true population variance by the factor of $\frac{n}{n-1}$. After the correction, we get a desired unbiased estimator for the population variance which we call a sample variance.

Example 4.5.5

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

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

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

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

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

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

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

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

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

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

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

Note that if the 20 students were the entire population then we would use μ instead of $\bar{x}$, σ^2 instead of s^2, and σ instead of s. In such a case, the **population variance**, σ^2, is equal to the sum of the last column (9.7375) divided by the total number of data values (20):

$$\sigma^2 = \frac{9.7375}{20} = 0.486875$$

and the **population standard deviation** (\sigma\) is equal to the square root of the population variance:

$$\sigma = \sqrt{0.486875} = 0.69776$$

Interactive Exercise 4.5.5.1

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

View this Interactive Content online

Interactive Exercise 4.5.5.2

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

View this Interactive Content online

Note

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION such as this one:

Descriptive Statistics Calculator

Regardless of the tool that you use, you still need to be aware of the context and use the appropriate notation for standard deviation σ or s.

Interactive Exercise 4.5.5.3

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

View this Interactive Content online

Interpreting the Mean and Standard Deviation Together

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

Chebyshev's Rule

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

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.

The Empirical Rule

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

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.

The empirical rule is also known as the 68-95-99.7 rule. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

Example 4.5.6

Suppose x is from a population with mean 50 and standard deviation 6 with bell-shape distribution.

- About 68% of the *x* values lie within one standard deviation of the mean. Therefore, about 68% of the *x* values lie between –1σ = (–1)(6) = –6 and 1σ = (1)(6) = 6 of the mean 50. The values 50 – 6 = 44 and 50 + 6 = 56 are within one standard deviation from the mean 50.
- About 95% of the x values lie within two standard deviations of the mean. Therefore, about 95% of the x values lie between –2σ = (–2)(6) = –12 and 2σ = (2)(6) = 12. The values 50 – 12 = 38 and 50 + 12 = 62 are within two standard deviations from the mean 50.
- About 99.7% of the x values lie within three standard deviations of the mean. Therefore, about 99.7% of the x values lie between –3σ = (–3)(6) = –18 and 3σ = (3)(6) = 18 from the mean 50. The values 50 – 18 = 32 and 50 + 18 = 68 are within three standard deviations of the mean 50.

Exercise 4.5.6

The population of scores on a college entrance exam have an approximate bell-shape distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68% of the y values lie between what two values? These values are _____.
b. About 95% of the y values lie between what two values? These values are _____.
c. About 99.7% of the y values lie between what two values? These values are _____.

Answer available at the end of this section.

Interactive Exercise 4.5.6

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

View this Interactive Content online

Note

It is important to note that the Empirical Rule only applies when the shape of the distribution of the data is bell-shaped and symmetric thus allowing us to sketch the following shape of the distribution based only on the two numbers: the mean and the standard deviation:

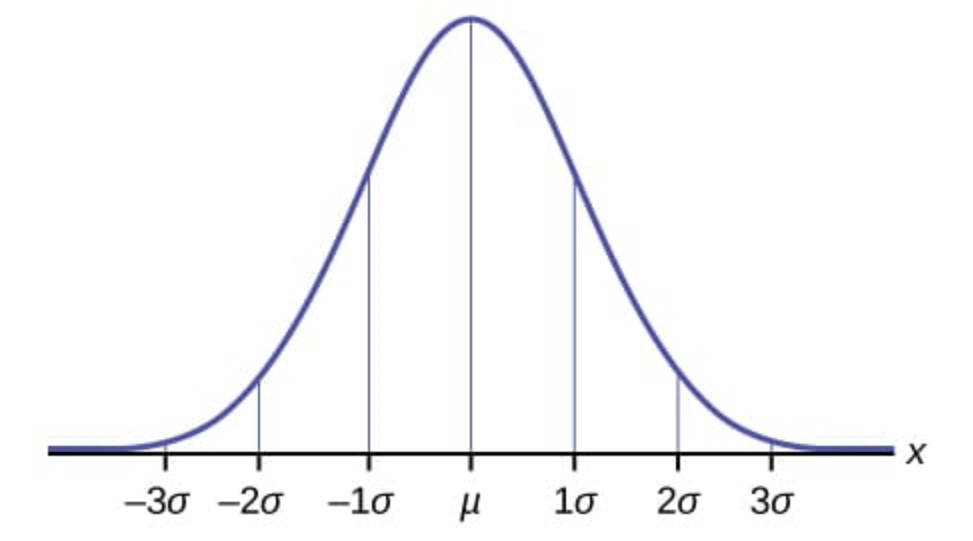


Figure 4.5.A

Interactive Exercise 4.5.7

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

View this Interactive Content online

Answers

Exercise 4.5.6

a. About 68% of the values lie between the values 41 and 63.
b. About 95% of the values lie between the values 30 and 74.
c. About 99.7% of the values lie between the values 19 and 85.

This page titled 4.5: Measures of the Center and Variation of the Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.
- zfmyzonep6 has no license indicated.
- 6.2: Null spaces by Isaiah Lankham, Bruno Nachtergaele, and Anne Schilling has no license indicated.

4.6: Measures of the Location of the Data and Outliers

Learning Objectives

- Recognize, describe, and calculate the measures of position.
- Identify an outlier.

In this section, we introduce the measures of relative standing, the purpose of which is to tell us how each individual value compares against the other values within the same data set. The easiest way to measure the relative standing of each observation is to **rank** them from the smallest to the largest. However, this is easier said than done for large data sets, so we will focus on the other two ways to measure the relative standing for each observation: the **z-score** and the **percentile**.

Z-Scores

Definition: z-score

Consider a population with known mean μ and known standard deviation σ, then the *z*-score of some observation x is:

$$z = \frac{x - \mu}{\sigma} \tag{4.6.1}$$

The term **standard score** is often used instead of z-score. A negative z-score indicates that the observation is below (less than) the mean, whereas a positive z-score indicates that the observation is above (greater than) the mean.

The z-score of an observation, therefore, can be used as a rough measure of its relative standing among all the observations comprising a data set. For instance, a z-score of 3 or more indicates that the observation is larger than most of the other observations; a z-score of –3 or less indicates that the observation is smaller than most of the other observations; a z-score near 0 indicates that the observation is located near the mean.

Example 4.6.1

If the mean of a normal distribution is five and the standard deviation is two, find the z-score of the value 11.

Solution

The calculation is as follows:

$$z = \frac{11 - 5}{2} = 3$$

Exercise 4.6.1

What is the z-score of $x = 1$ in the data set with mean 12 and standard deviation 3?

Answer available at the end of this section.

Interactive Exercise 4.6.1

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

View this Interactive Content online

If two distributions have the same shape or, more generally, if they differ only by center and variation, then z-scores can be used to compare the relative standings of two observations from those distributions.

Example 4.6.2

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

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

Solution

For Angie:

$$z = \frac{26.2 - 27.2}{0.8} = -1.25$$

For Beth:

$$z = \frac{27.3 - 30.1}{1.4} = -2$$

Since Beth's z-score is smaller than Angie's z-score, we conclude that Beth has a relatively faster time when compared to her team.

Exercise 4.6.2

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

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

Answer available at the end of this section.

Interactive Exercise 4.6.2

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

View this Interactive Content online

Z-scores and the Empirical Rule

Recall that if a variable has a bell-shape distribution with mean μ and standard deviation σ, then the **Empirical Rule** says the following:

- About 68% of the values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).
- About 95% of the values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the values lie between –3σ and +3σ of the mean μ (within three standard deviations of the mean).

Also, note that

- The z-scores for +1σ and –1σ are +1 and –1, respectively.
- The z-scores for +2σ and –2σ are +2 and –2, respectively.
- The z-scores for +3σ and –3σ are +3 and –3 respectively.

So that the Empirical Rule says the following:

- About 68% of the values have z-scores between –1 and +1.
- About 95% of the values have z-scores between –2 and +2.
- About 99.7% of the values have z-scores between –3 and +3.

Definition: Outlier

An **outlier** is an observation that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Using z-scores to identify outliers

Since most of the observations in any dataset have the z-score between -3 and 3, any observation with the z-score less than -3 or more than 3 is considered to be an outlier as it very unlikely.

Example 4.6.3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6 cm. Suppose a male from Chile was 158 cm tall from 2009 to 2010. Is he an outlier?

Solution

The z-score of 158 is

$$z = \frac{158 - 170}{6} = \frac{-12}{6} = -2$$

Since -2 is between -3 and 3, the male is not an outlier.

Exercise 4.6.3

Jerome averages 18 points a game with a standard deviation of 3 points. Suppose Jerome scores 28 points in one game. Is this result an outlier?

Answer available at the end of this section.

Interactive Exercise 4.6.3

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

View this Interactive Content online

Percentiles

Another way to measure the relative standing is computing the **percentiles**. The p-th **percentile** is a value that separates the bottom p% of the data form the top (100-p)% of the data. In other words, the percentile indicates the relative standing of an observation when data are sorted from smallest to largest. For example, 15% of data values are less than or equal to the 15th percentile.

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

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a

certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies. Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

Note

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

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

Also, note that to score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Example 4.6.4

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

Solution

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

Exercise 4.6.4

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

Answer available at the end of this section.

Interactive Exercise 4.6.4

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

View this Interactive Content online



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

To find the percentile of a particular observation, we use the following procedure:

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

It is common to define x as the the number of data values counting from the bottom of the data list up to **and** including the data value. This may slightly change the result of the computation. However, percentiles are mostly used with very large populations, thus, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

Example 4.6.5

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

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

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution

a. Counting from the bottom of the list, there are 18 data values less than 58.

   $x = 18$ and $n = 29$ thus $\frac{x}{n}(100) = \frac{18}{29}(100) = 62.07$. 58 is the 62nd percentile.

b. Counting from the bottom of the list, there are three data values less than 25.

   $x = 3$ and $n = 29$. $\frac{x}{n}(100) = \frac{3}{29}(100) = 10.34$. 25 is the 10th percentile.

Exercise 4.6.5

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

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

Find the percentiles for 47 and 31.

Answer available at the end of this section.

Interactive Exercise 4.6.5

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

View this Interactive Content online

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

A Formula for Finding the p-th Percentile

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

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

1. Order the data from smallest to largest.
2. Calculate $i = \frac{k}{100}(n+1)$
3. If i is an integer, then the kth percentile is the data value in the ith position in the ordered set of data. If i is not an integer, then round i up and round i down to the nearest integers. The p-th percentile is the average the two data values in these two positions in the ordered data set.

Example 4.6.6

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

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

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution

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

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

b.
- $k =$ 83rd percentile
- $i = theindex$
- $n = 29$

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

Exercise 4.6.6

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

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

Calculate the 20th percentile and the 55th percentile.

Answer

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

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

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The second quartile, Q_2, is also known as the **median**, *M*, and is the same as the 50th percentile. We can also extend the definition, and define the zeroth quartile, Q_0, as the 0th percentile (the smallest value), and the forth quartile, Q_4, as the 100th percentile (the largest value).

Example 4.6.7

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

Solution

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

Exercise 4.6.7

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

Answer available at the end of this section.

Interactive Exercise 4.6.7

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

View this Interactive Content online

To calculate quartiles, the data must be ordered from smallest to largest. The median divides ordered data into halves that is half the values are the same number or smaller than the median, and half the values are the same number or larger. Quartiles divide ordered data into quarters, that is, Q1 divides the bottom half into two halves and Q3 divides the top half into two halves.

Example 4.6.8

Consider the following data.

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

Ordered from smallest to largest:

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

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

$$\frac{6.8 + 7.2}{2} = 7 \tag{4.6.2}$$

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

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

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

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

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

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

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

The **third quartile**, Q3, is nine. Three-fourths (75%) of the ordered data set is less than nine. One-fourth (25%) of the ordered data set is greater than nine. The third quartile is part of the data set in this example.

Exercise 4.6.8

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

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes; 10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes; 30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the five-value summary.

Answer available at the end of this section.

Interactive Exercise 4.6.8

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

View this Interactive Content online

Definition: Interquartile Range

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

$$IQR = Q_3 - Q_1 \tag{2.4.1}$$

Using IQR to identify outliers

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

Example 4.6.9

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

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

Answer

Order the data from smallest to largest.

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

$$M = 488,800$$

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

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

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

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

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

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

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

Exercise 4.6.9

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

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

Answer

Order the data from smallest to largest.

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

Median = $54,000

$$Q_1 = \$40,500$$

$$Q_3 = \$69,000$$

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

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

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

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

No salary is less than –$2,250. However, $120,000 is more than $11,750, so $120,000 is a potential outlier.

Exercise 4.6.9

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

View this Interactive Content online

Percentile or Z-score?

Which one is better? The use of z-scores as a measure of relative standing can be refined and made more precise by applying the Chebyshev's rule, but percentiles usually give a more accurate and meaningful measure of relative standing than z-scores. However, to compute the percentile one needs to have access to the entire data set while to compute the z-score one needs to know only the mean and standard deviation of the data set. With very little information, z-scores provide a feasible alternative to percentiles for measuring relative standing. For example, for a student to be able to compute the percentile of their exam score the instructor must allow the access to the entire list of scores which is impossible, but to compute the z-score the instructor must share only the mean and the standard deviation which is something instructors are more comfortable doing.

Answers

Exercise 4.6.1

$z = \frac{1-12}{3} \approx -3.67$

Exercise 4.6.2

For Fredo:

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

For Karl:

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

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

Exercise 4.6.3

The z-score of 28 is

$$z = \frac{28-18}{3} = \frac{10}{3} = 3.33$$

Since 3.33 is greater than 3, this result is an outlier.

Exercise 4.6.4

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

Exercise 4.6.5

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

$x = 15$ and $n = 30$. $\frac{x}{n}(100) = \frac{15}{30}(100) = 50$. 47 is the 50th percentile.

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

$x = 8$ and $y = 30$. $\frac{x}{n}(100) = \frac{8}{30}(100) = 26.67$. 31 is the 27th percentile.

Exercise 4.6.7

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

Exercise 4.6.8

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

This page titled 4.6: Measures of the Location of the Data and Outliers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 1.1: Real Numbers - Algebra Essentials by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.
- 6.2: Null spaces by Isaiah Lankham, Bruno Nachtergaele, and Anne Schilling has no license indicated.
- ghjmtdzi|ellaqld has no license indicated.

5: PROBABILITY

Close message

Learn more

Welcome to your new guide!

Click on "New" to build out your guide.

- Topic hierarchy

5.1: Introduction on Probability

5.2: Expected Value

5.3: Probability Rules- “And” and “Or”

5.4: Counting Methods

5.5: Conditional Probabilities

This page titled 5: Probability is shared under a license and was authored, remixed, and/or curated by LibreTexts.

5.1: Introduction on Probability

Learning Objectives

Students will be able to:

- Understand and use probability terminology correctly.
- Calculate the probability of a given event for equally likely outcomes.
- Determine the compliment of a given event and find its probability.

Toss a thumb tack one time. Do you think the tack will land with the point up or the point down?

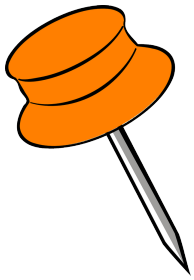


Figure 5.1.1: Which Way Will a Tack Fall? (Thumbtack, n.d.)

We cannot predict which way the tack will land before we toss it. Sometimes it will land with the point up and other times it will land with the point down. Tossing a tack is a random experiment since we cannot predict what the outcome will be. We do know that there are only two possible outcomes for each trial of the experiment: lands points up or lands point down. If we repeat the experiment of tossing the tack many times we might be able to guess how likely it is that the tack will land point up.

Definitions

- A **random experiment** is an activity or an observation whose outcome cannot be predicted ahead of time.
- The result of an experiment is called an **outcome**.
- A **trial** is one repetition of a random experiment.
- The **sample space** is the set of all possible outcomes for a random experiment.
- An **event** is a subset of the sample space. An **event** can be thought of as a collection of outcomes that are grouped together.
- A **simple event** is an event that consists of only one outcome. For example, rolling a die and getting a "2" is a simple event.
- The **probability** of an event is the measure of how likely it will happen. It can be expressed as a number between 0 and 1, or as a percentage between 0% and 100% The probability of an **impossible event** is 0 (or 0%), and the probability of a **certain event** is 1 (or 100%).

Do you think chances of the tack landing point up and the tack landing point down are the same? This is an example where your intuition may be wrong. Having only two possible outcomes does not mean each outcome has a 50/50 chance of happening. In fact we are going to see that the probability of the tack landing point up is about 66%.

To begin to answer this question, toss a tack 10 times. For each toss record whether the tack lands point up or point down.

Table 5.1.2: Toss a Tack Ten Times

Trial	1	2	3	4	5	6	7	8	9	10
Up/Down	Up	Down	Up	Up	Down	Up	Up	Down	Up	Down

The random experiment here is tossing the tack one time. The possible outcomes for the experiment are that the tack lands point up or the tack lands point down so the sample space is S = {point up, point down}. We are interested in the event E that the tack lands point up, E = {point up}.

Based on our data we would say that the tack landed point up six out of 10 times. The fraction, $\frac{6}{10}$, is called the relative frequency. Since $\frac{6}{10} = 0.60$ we would guess that probability of the tack landing point up is about 60%.

Let's repeat the experiment by tossing the tack ten more times.

Table 5.1.3: Toss a Tack Ten More Times

Trial	1	2	3	4	5	6	7	8	9	10
Up/Down	Up	Up	Up	Down	Down	Up	Up	Down	Up	Up

This time the tack landed point up seven out of 10 times or 70% of the time. If we tossed the tack another 10 times we might get a different result again. The probability of the tack landing point up refers to what happens when we toss the tack many, many times. Let's toss the tack 150 times and count the number of times it lands point up. Along the way we will look at the proportion of landing point up.

Table 5.1.4: Toss a Tack Many Times

Trials	Number Up	Total Number of Up	Total Number of Trials	Proportion
0-10	6	6	10	6/10=0.60
11-20	7	13	20	13/20=0.65
21-30	8	21	30	21/30=0.70
31-40	6	27	40	27/40=0.68
41-50	8	35	50	35/50=0.70
51-60	8	43	60	43/60=0.72

Trials	Number Up	Total Number of Up	Total Number of Trials	Proportion
61-70	4	47	70	47/70=0.67
71-80	7	54	80	54/80=0.68
81-90	5	59	90	59/90=0.66
91-100	6	65	100	65/100=0.65
101-110	10	75	110	75/110=0.68
111-120	5	80	120	80/120=0.67
121-130	8	88	130	88/130=0.68
131-140	4	92	140	92/140=0.66
141-150	7	99	150	99/150=0.66

When we have a small number of trials the proportion varies quite a bit. As we start to have more trials the proportion still varies but not by as much. It appears that the proportion is around 0.66 or 66%. We would have to do about 100,000 trials to get a better approximation of the actual probability of the tack landing point up.

The tack landed point up 99 out of 150 trials. The probability *P* of event *E* is written as:

$$P(E) = \frac{\#\text{ of trials with point up}}{\text{total number of trials}} = \frac{99}{150} \approx 0.66 \tag{5.1.1}$$

We would say the probability that the tack lands point up is about 66%.

Equally Likely Outcomes

In some experiments all the outcomes have the same chance of happening. If we roll a fair die the chances are the same for rolling a two or rolling a five. If we draw a single card from a well shuffled deck of cards, each card has the same chance of being selected. We call outcomes like these equally likely. Drawing names from a hat or drawing straws are other examples of equally likely outcomes. The tack tossing example did not have equally likely outcomes since the probability of the tack landing point up is different than the probability of the tack landing point down.

Definitions

An experiment has **equally likely outcomes** if every outcome has the same probability of occurring.

For equally likely outcomes, the **probability of event A**, *P(A)*, is:

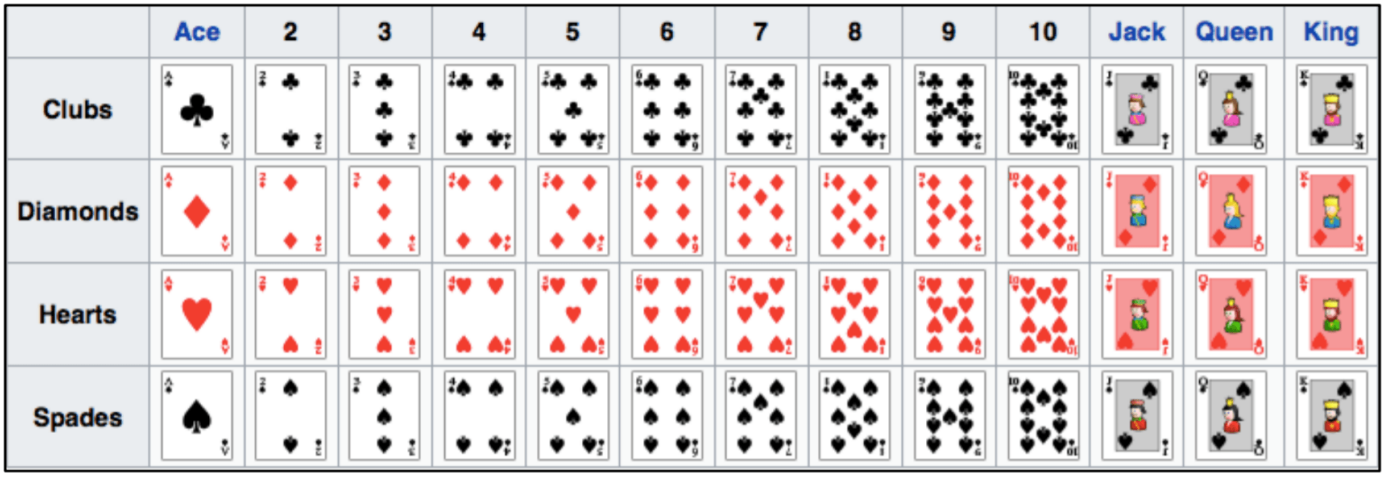
$$P(A) = \frac{\text{number of ways for A to occur}}{\text{total number of outcomes}}. \tag{5.1.2}$$

Round Off Rule: Give probabilities as a fraction or as a decimal number rounded to three decimal places.

Definition: Cards

A standard deck of 52 playing cards consists of four **suits** (hearts, spades, diamonds and clubs). Spades and clubs are black while hearts and diamonds are red. Each suit contains 13 cards, each of a different **rank**: An Ace (which in many games functions as both a low card and a high card), cards numbered 2 through 10, a Jack, a Queen, and a King.

The image below gives an example of a complete deck of 52 cards.



Example 5.1.1: Simple Probabilities with Cards

Draw a single card from a well shuffled deck of 52 cards. Each card has the same chance of being drawn so we have equally likely outcomes. Find following probabilities of the given events.

1. *P*(card is red)

P(card is red) = $\frac{\text{number of red cards}}{\text{total number of cards}} = \frac{26}{52} = \frac{1}{2}$

The probability that the card is red is $\frac{1}{2}$.

2. *P*(card is a heart)

P(card is a heart) = $\frac{\text{number of hearts}}{\text{total number of cards}} = \frac{13}{52} = \frac{1}{4}$

The probability that the card is a heart is $\frac{1}{4}$.

3. *P*(card is a red 5)

P(card is a red 5) = $\frac{\text{number of red fives}}{\text{total number of cards}} = \frac{2}{52} = \frac{1}{26}$

The probability that the card is a red five is $\frac{1}{26}$.

Example 5.1.2: Simple Probabilities with a Fair Die

Roll a fair die one time. The sample space is S = {1, 2, 3, 4, 5, 6}. Find the following probabilities of the given events.

1. *P*(roll a four)

$$P(\text{roll a four}) = \frac{\text{number of ways to roll a four}}{\text{total number of ways to roll a die}} = \frac{1}{6}$$

The probability of rolling a four is $\frac{1}{6}$.

2. *P*(roll an odd number)

The event roll an odd number is E = {1, 3, 5}.

$$P(\text{roll an odd number}) = \frac{\text{number of ways to roll an odd number}}{\text{total number of ways to roll a die}} = \frac{3}{6} = \frac{1}{2}$$

The probability of rolling an odd number is $\frac{1}{2}$.

3. *P*(roll a number less than five)

The event roll a number less than five is F = {1, 2, 3, 4}.

$$P(\text{roll a number less than five}) = \frac{\text{number of ways to roll number less than five}}{\text{total number of ways to roll a die}} = \frac{4}{6} = \frac{2}{3}$$

The probability of rolling a number less than five is $\frac{2}{3}$.

Example 5.1.3: Simple Probability with Books

A small bookcase contains five math, three English and seven science books. A book is chosen at random. What is the probability that a math book is chosen?



Figure 5.1.6: Books on a Shelf. (Bookshelf, 2011)

Solution

Since the book is chosen at random each book has the same chance of being chosen and we have equally likely outcomes.

$$P(\text{ math book }) = \frac{\text{number of ways to choose a math book}}{\text{total number of books}} = \frac{5}{15} = \frac{1}{3} \tag{5.1.3}$$

The probability a math book was chosen is $\frac{1}{3}$.

Three Ways of Finding Probabilities

There are three ways to find probabilities. In the tack tossing example we calculated the probability of the tack landing point up by doing an experiment and recording the outcomes. This was an example of an empirical probability. The probability of getting a red jack in a card game or rolling a five with a fair die can be calculated from mathematical formulas. These are examples of theoretical probabilities. The third type of probability is a subjective probability. Saying that there is an 80% chance that you will go to the beach this weekend is a subjective probability. It is based on experience or guessing.

Definitions

- A **theoretical probability** is based on a mathematical model where all outcomes are equally likely to occur.
- An **empirical probability** is based on an experiment or observation and is the relative frequency of the event occurring.
- A **subjective probability** is an estimate (a guess) based on experience or intuition.

Law of Large Numbers

When studying probabilities, many times the law of large numbers will apply. If you want to observe what the probability is of getting tails up when flipping a coin, you could do an experiment. Suppose you flip a coin 20 times and the coin comes up tails nine times. Then, using an empirical probability, the probability of getting tails is 9/20 = 45%. However, we know that the theoretical probability for getting tails should be 1/2 = 50%. Why is this different? It is because there is error inherent to sampling methods. However, if you flip the coin 100 times or 1000 times, and use the information to calculate an empirical probability for getting tails up, then the probabilities you will observe will become closer to the theoretical probability of 50%. This is the law of large numbers.

Definition: Law of Large Numbers

The **law of large numbers** means that with larger numbers of trials of an experiment the observed empirical probability of an event will approach the calculated theoretical probability of the same event.

Complements

If there is a 75% chance of rain today, what are the chances it will not rain? We know that there are only two possibilities. It will either rain or it will not rain. Because the sum of the probabilities for all the outcomes in the sample space must be 100% or 1.00, we know that

$$P(\text{will rain}) + P(\text{will not rain}) = 100\%.$$

Rearranging this we see that

$$P(\text{will not rain}) = 100\% - P(\text{will rain}) = 100\% - 75\% = 25\%.$$

The events E = {will rain} and F = {will not rain} are called complements.

Definition: Complement

The **complement** of event *E*, denoted by $\overline{E}$, is the set of outcomes in the sample space that are not in the event *E*. The probability of $\overline{E}$ is given by $P(\overline{E}) = 1 - P(E)$.

Example 5.1.4: Complements with Cards

Draw a single card from a well shuffled deck of 52 cards.

1. Look at the suit of the card. Here the sample space S = {spades, clubs, hearts, diamonds}. If event *E* = {spades} the complement $\overline{E}$ = {clubs, hearts, diamonds}.
2. Look at the value of the cards. Here the sample space is S = {A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K}. If the event *E* = {the number is less than 7} = {A, 2, 3, 4, 5, 6} the complement $\overline{E}$ = {7, 8, 9, 10, J, Q, K}.

Example 5.1.5: Complements with Trains

A train arrives on time at a particular station 85% of the time. Does this mean that the train is late 15% of the time? The answer is no. The complement of *E* = {on time} is not $\overline{E}$ = {late}. There is a third possibility. The train could be early. The sample space is S = {on time, early, late} so the complement of *E* = {on time} is $\overline{E}$ = {early or late}. Based on the given information we cannot find *P*(late) but we can find *P*(early or late) = 15%.

Impossible Events and Certain Events

Recall that $P(A) = \frac{\text{number of ways for A to occur}}{\text{total number of outcomes}}$. What does it mean if we say the probability of the event is zero? $P(A) = \frac{\text{number of ways for A to occur}}{\text{total number of outcomes}} = 0$. The only way for a fraction to equal zero is when the numerator is zero. This means there is no way for event *A* to occur. A probability of zero means that the event is impossible.

What does it mean if we say the probability of an event is one? $P(A) = \frac{\text{number of ways for A to occur}}{\text{total number of outcomes}} = 1$. The only way for a fraction to equal one is if the numerator and denominator are the same. The number of ways for *A* to occur is the same as the number of outcomes. There are no outcomes where *A* does not occur. *A* probability of 1 means that the event always happens.

- $P(A) = 0$ means that A is impossible
- $P(A) = 1$ means that A is certain

This page titled 5.1: Introduction on Probability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier via source content that was edited to the style and standards of the LibreTexts platform.

- 3.1: Basic Probabilities and Probability Distributions; Three Ways to Define Probabilities by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.
- 10.1: Basic Concepts by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf.
- 3.4: Expected Value and Law of Large Numbers by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.

5.2: Expected Value

Learning Objectives

Students will be able to:

- Identify valid probability distributions.
- Compute the expected value from a probability distribution or an experiment.
- Understand the implications of the Gamblers' Fallacy and how it related to streaks.

Probability Distributions

A probability distribution (probability space) is a sample space paired with the probabilities for each outcome in the sample space. If we toss a fair coin and see which side lands up, there are two outcomes, heads and tails. Since the coin is fair these are equally likely outcomes and have the same probabilities. The probability distribution would be *P*(heads) = 1/2 and *P*(tails) = 1/2. This is often written in table form:

Table 5.2.1: Probability Distribution for a Fair Coin

Outcome	Heads	Tails
Probability	1/2	1/2

A **probability distribution** for an experiment is a list of all the possible outcomes and their corresponding probabilities.

In probability problems when we roll two dice, it is helpful to think of the dice as being different colors. Let's assume that one die is red and the other die is green. We consider getting a three on the red die and a five on the green die different than getting a five on the red die and a three on the green die. In other words, when we list the outcomes the order matters. The possible outcomes of rolling two dice and looking at the sum are given in Table 5.2.2.

Table 5.2.2: All Possible Sums of Two Dice

1+1 = 2	1+2 = 3	1+3 = 4	1+4 = 5	1+5 = 6	1+6 = 7
2+1 = 3	2+2 = 4	2+3 = 5	2+4 = 6	2+5 = 7	2+6 = 8
3+1 = 4	3+2 = 5	3+3 = 6	3+4 = 7	3+5 = 8	3+6 = 9
4+1 = 5	4+2 = 6	4+3 = 7	4+4 = 8	4+5 = 9	4+6 = 10
5+1 = 6	5+2 = 7	5+3 = 8	5+4 = 9	5+5 = 10	5+6 = 11

6+1 = 7	6+2 = 8	6+3 = 9	6+4 = 10	6+5 = 11	6+6 = 12

Table 5.2.3: Probability Distribution for the Sum of Two Fair Dice

Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
Reduced Probability	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Example 5.2.1: Valid and Invalid Probability Distributions

Are the following valid probability distributions?

Table 5.2.4

Outcome	A	B	C	D	E
Probability	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

This is a valid probability distribution. All the probabilities are between zero and one inclusive and the sum of the probabilities is 1.00.

Table 5.2.5:

Outcome	A	B	C	D	E	F
Probability	0.45	0.80	-0.20	-0.35	0.10	0.20

This is not a valid probability distribution. The sum of the probabilities is 1.00, but some of the probabilities are not between zero and one, inclusive.

Table 5.2.6:

Outcome	A	B	C	D
Probability	0.30	0.20	0.40	0.25

This is not a valid probability distribution. All of the probabilities are between zero and one, inclusive, but the sum of the probabilities is 1.15 not 1.00.

Would you buy a lottery ticket with the numbers 1, 2, 3, 4, 5? Do you think that a winning ticket with five consecutive numbers is less likely than a winning ticket with the numbers 2, 14, 18, 23 and 32? If you are playing a slot machine in Las Vegas and you have lost the last 10 times, do you keep playing the same machine because you are "due for a win?" Have you ever wondered how a casino can afford to offer meals and rooms at such cheap rates? Should you play a game of chance at a carnival? How much should an organization charge for raffle tickets for their next fund raiser? All of these questions can be answered using probabilities.

Definition: Expected Value

Suppose the there are n outcomes for a given experiment where the values of the outcomes are given by x where x can take on the n values $x_1, x_2, x_3, \ldots, x_n$. If the probability that each of these values occurs is $p_1, p_2, p_3, \ldots, p_n$, respectively, then the **expected value** of the outcomes is

$$E(x) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots + x_np_n \tag{5.2.1}$$

The **expected value** is the average gain or loss of an event if the experiment is repeated many times.

We can use the probability distribution table to compute the expected value by multiplying each outcome by the probability of that outcome, then adding up the products.

Example 5.2.2: Expected Value for Raffle Tickets

Valley View Elementary is trying to raise money to buy tablets for their classrooms. The PTA sells 2000 raffle tickets at $3 each. First prize is a flat-screen TV worth $500. Second prize is an android tablet worth $375. Third prize is an e-reader worth $200. Five $25 gift certificates will also be awarded. What are the expected winnings for a person who buys one ticket?

Solution

We need to write out the probability distribution before we find the expected value.

A total of eight tickets are winners and the other 1992 tickets are losers.

Table 5.2.7: Probability Distribution for the Valley View Raffle

Outcome	Win $500	Win $375	Win $200	Win $25	Win $0
Probability	$\frac{1}{2000}$	$\frac{1}{2000}$	$\frac{1}{2000}$	$\frac{5}{2000} = \frac{1}{400}$	$\frac{1992}{2000} = \frac{249}{250}$

Now use the formula for the expected value (Equation 5.2.1).

$$E = \$500\left(\frac{1}{2000}\right) + \$375\left(\frac{1}{2000}\right) + \$200\left(\frac{1}{2000}\right) + \$25\left(\frac{1}{400}\right) + \$0\left(\frac{249}{250}\right) \\ = 0.60 \tag{5.2.2}$$

It costs $3 to buy a ticket but we only win an average of $0.60 per ticket. That means the expected winnings per ticket are $0.60 - $3 = -$2.40.

We would expect to lose an average of $2.40 for each ticket bought. This means that the school will earn an average of $2.40 for each ticket bought for a profit of $2.40 · 2000 = $4800.

In general, if the expected value of a game is negative, it is not a good idea to play the game, since on average 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. If the expected value of a game is 0, we call it a **fair game**, since neither side has an advantage.

Not surprisingly, the expected value for casino games is negative for the player, which is positive for the casino. It must be positive or they would go out of business. Players just need to keep in mind that when they play a game repeatedly, their expected value is negative. That is fine so long as you enjoy playing the game and think it is worth the cost. But it would be wrong to expect to come out ahead.

Example 5.2.3: Expected Value for Profit from a Purchase

A real estate investor buys a parcel of land for $150,000. He estimates the probability that he can sell it for $200,000 to be 0.40, the probability that he can sell it for $160,000 to be 0.45 and the probability that he can sell it for $125,000 to be 0.15. What is the expected profit for this purchase?

Solution

Find the profit for each situation first. $200,000 – $150,000 = $50,000 profit, $160,000 - $150,000 = $10,000 profit, and $125,000 - $150,000 = -$25,000 profit (loss).

The probability distribution is

Table 5.2.8: Probability Distribution for a Real Estate Purchase

Outcome	$50,000	$10,000	-$25,000
Probability	0.40	0.45	0.15

$$E = 50,000(0.40) + 10,000(0.45) + (-25,000)(0.15) \\ = \$20,750 \tag{5.2.3}$$

The expected profit from the purchase is $20,750.

Example 5.2.4: Expected Value for Life Insurance

The cost of a $50,000 life insurance policy is $150 per year for a person who is 21-years old. Assume the probability that a person will die at age 21 is 0.001. What is the company's expected profit if the company sells 10,000 policies to 21-year olds?

Solution

There are two outcomes. If the person lives the insurance company makes a profit of $150. The probability that the person lives is 1-0.001=0.999. If the person dies the company takes in $150 and pays out $50,000 for a loss of $49,850.

Table 5.2.9: Probability Distribution for Life Insurance

Outcome	$150	-$49,850
Probability	0.999	0.001

The expected value for one policy is:

$$E(x) = (\$150)(0.999) + (-\$49,850)(0.001) = \$149.80$$

If the company sells 10,000 policies at a profit of $149.80 each, the total expected profit is $\$149.80(10,000) = \$1,498,000$.

Gamblers' Fallacy and Streaks

Often times a gambler on a losing streak will keep betting in the belief that his/her luck must soon change. Consider flipping a fair coin. Each toss of the coin is independent of all the other tosses, meaning the past outcomes do not influence the probability of getting a heads or tails on the next coin toss. Assume the coin has landed heads up the last eight times. Some people erroneously believe that the coin is more likely to land tails up on the next toss. In reality, the coin still has a 50% chance of landing tails up. It does not matter what happened the last eight tosses.

Definition: Gambler's Fallacy

The **gambler's fallacy** is the mistaken belief that a streak of losses makes a person due for a streak of wins and vice versa.

Example 5.2.5: Streaks

Toss a fair coin seven times and record which side lands up. For example HHTHTTT would represent getting a head on the first, second, and fourth tosses and a tail on the other tosses. Is a streak of all heads less likely than the other possible outcomes?

Solution

As we will see in Section 3.5, there are 128 possible ways to toss a coin seven times. Some of the possibilities are $HHTTHHT$, $HTHTHTH$, $HHHHTTT$, and $HTTHTTH$. Because tossing coins are independent events and the coin is fair each of these 128 possibilities has the same probability.

$$P(\text{HHTTHHT}) = \frac{1}{128}$$

$$P(\text{HTHTHTH}) = \frac{1}{128}$$

etc.

This also means that the probability of getting all heads is

$$P(\text{HHHHHHH}) = \frac{1}{128}.$$

> Getting a streak of all heads has exactly the same probability as any other possible outcome.

This page titled 5.2: Expected Value is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier via source content that was edited to the style and standards of the LibreTexts platform.

- 3.1: Basic Probabilities and Probability Distributions; Three Ways to Define Probabilities by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.
- 3.4: Expected Value and Law of Large Numbers by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.
- 10.5: Expected Value by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: https://www.sccollege.edu/OER/Documents/MATH 105/Math For Liberal Art Students (2017).pdf.

5.3: Probability Rules- “And” and “Or”

Learning Objectives

Students will be able to:

- Determine if two events are mutually exclusive and/or independent.
- Apply the "Or" rule to calculate the probability that either of two events occurs.
- Apply the "And" rule to calculate the probability that both of two events occurs.

Many probabilities in real life involve more than one event. If we draw a single card from a deck we might want to know the probability that it is either red or a jack. If we look at a group of students, we might want to know the probability that a single student has brown hair and blue eyes. When we combine two events we make a single event called a **compound event**. To create a compound event, we can use the word “and” or the word “or” to combine events. It is very important in probability to pay attention to the words “and” and “or” if they appear in a problem. The word “and” restricts the field of possible outcomes to only those outcomes that simultaneously describe all events. The word “or” broadens the field of possible outcomes to those that describe one or more events.

Example 5.3.1: Counting Students

Suppose a teacher wants to know the probability that a single student in her class of 30 students is taking either Art or English. She asks the class to raise their hands if they are taking Art and counts 13 hands. Then she asks the class to raise their hands if they are taking English and counts 21 hands. The teacher then calculates

$$P(\text{Art or English}) = \frac{13+21}{30} = \frac{33}{30}$$

The teacher knows that this is wrong because probabilities must be between zero and one, inclusive. After thinking about it she remembers that nine students are taking both Art and English. These students raised their hands each time she counted, so the teacher counted them twice. When we calculate probabilities we have to be careful to count each outcome only once.



Figure 5.3.1: College Classroom (Colwell, 2013)

Mutually Exclusive Events

An experiment consists of drawing one card from a well shuffled deck of 52 cards. Consider the events *E*: the card is red, *F*: the card is a five, and *G*: the card is a spade. It is possible for a card to be both red and a five at the same time but it is not possible for a card to be both red and a spade at the same time. It would be easy to accidentally count a red five twice by mistake. It is not possible to count a red spade twice.

Definition: Mutually Exclusive

Two events are **mutually exclusive** if they have no outcomes in common.

Example 5.3.2: Mutually Exclusive with Dice

Two fair dice are tossed and different events are recorded. Let the events *E*, *F* and *G* be as follows:

- *E* = {the sum is five} = {(1, 4), (2, 3), (3, 2), (4, 1)}
- *F* = {both numbers are even} = {(2, 2), (2, 4), (2, 6), (4, 2), (4, 4), (4, 6), (6, 2), (6, 4), (6, 6)}
- *G* = {both numbers are less than five} = {(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4), (4,1), (4, 2), (4, 3), (4,4)}

1. Are events *E* and *F* mutually exclusive?

Yes. *E* and *F* are mutually exclusive because they have no outcomes in common. It is not possible to add two even numbers to get a sum of five.

2. Are events *E* and *G* mutually exclusive?

No. *E* and *G* are not mutually exclusive because they have some outcomes in common. The pairs (1, 4), (2, 3), (3, 2) and (4, 1) all have sums of 5 and both numbers are less than five.

3. Are events *F* and *G* mutually exclusive?

No. *F* and *G* are not mutually exclusive because they have some outcomes in common. The pairs (2, 2), (2, 4), (4, 2) and (4, 4) all have two even numbers that are less than five.

Addition Rule for “Or” Probabilities

The addition rule for probabilities is used when the events are connected by the word “or”. Remember our teacher in Example 5.3.1 at the beginning of the section? She wanted to know the probability that her students were taking either art or English. Her problem was that she counted some students twice. She needed to add the number of students taking art to the number of students taking English and then subtract the number of students she counted twice. After dividing the result by the total number of students she will find the desired probability. The calculation is as follows:

$$\begin{aligned} P(\text{art or English}) &= \frac{\#\text{ taking art} + \#\text{ taking English - }\#\text{ taking both}}{\text{total number of students}} \\ &= \frac{13+21-9}{30} \\ &= \frac{25}{30} \approx 0.833 \end{aligned}$$

The probability that a student is taking art or English is 0.833 or 83.3%.

When we calculate the probability for compound events connected by the word “or” we need to be careful not to count the same thing twice. If we want the probability of drawing a red card or a five we cannot count the red fives twice. If we want the probability a person is blonde-haired or blue-eyed we cannot count the blue-eyed blondes twice. The addition rule for probabilities adds the

number of blonde-haired people to the number of blue-eyed people then subtracts the number of people we counted twice.

Addition Rule for "Or" Probabilities

If *A* and *B* are any events then

$$P(A \text{ or } B) = P(A) + P(B) – P(A \text{ and } B). \tag{5.3.1}$$

If *A* and *B* are mutually exclusive events then $P(A \text{ and } B) = 0$, so then

$$P(A \text{ or } B) = P(A) + P(B). \tag{5.3.2}$$

Example 5.3.3: Additional Rule for Drawing Cards

A single card is drawn from a well shuffled deck of 52 cards. Find the probability that the card is a club or a face card.

Solution

There are 13 cards that are clubs, 12 face cards (J, Q, K in each suit) and 3 face cards that are clubs.

$$\begin{aligned} P(\text{club or face card}) &= P(\text{club}) + P(\text{face card}) - P(\text{club and face card}) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= \frac{22}{52} = \frac{11}{26} \approx 0.423 \end{aligned}$$

The probability that the card is a club or a face card is approximately 0.423 or 42.3%.

A simple way to check this answer is to take the 52 card deck and count the number of physical cards that are either clubs or face cards. If you were to set aside all of the clubs and face cards in the deck, you would end up with the following:

{2 Clubs, 3 Clubs, 4 Clubs, 5 Clubs, 6 Clubs, 7 Clubs, 8 Clubs, 9 Clubs, 10 Clubs, J Clubs, Q Clubs, K Clubs, A Clubs, J Hearts, Q Hearts, K Hearts, J Spades, Q Spades, K Spades, J Diamonds, Q Diamonds, K Diamonds}

That is 22 cards out of the 52 card deck, which gives us a probably of:

$$\frac{22}{52} = \frac{11}{26} \approx 0.423$$

This confirms our earlier answer using the formal Addition Rule.

Example 5.3.4: Addition Rule for Tossing a Coin and Rolling a Die

An experiment consists of tossing a coin then rolling a die. Find the probability that the coin lands heads up or the number is five.

Solution

Let H represent heads up and T represent tails up. The sample space for this experiment is S = {H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6}.

- There are six ways the coin can land heads up, {H1, H2, H3, H4, H5, H6}.
- There are two ways the die can land on five, {H5, T5}.
- There is one way for the coin to land heads up and the die to land on five, {H5}.

$$P(\text{heads or five}) = P(\text{heads}) + P(\text{five}) - P(\text{both heads and five})$$
$$= \frac{6}{12} + \frac{2}{12} - \frac{1}{12}$$
$$= \frac{7}{12} = \approx 0.583$$

The probability that the coin lands heads up or the number is five is approximately 0.583 or 58.3%.

Example 5.3.5: Addition Rule for Satisfaction of Car Buyers

Two hundred fifty people who recently purchased a car were questioned and the results are summarized in the following table.

Table 5.3.1: Satisfaction of Car Buyers

	Satisfied	Not Satisfied	Total
New Car	92	28	120
Used Car	83	47	130
Total	175	75	250

Find the probability that a person bought a new car or was not satisfied.

Solution

$$P(\text{new car or not satisfied}) = P(\text{new car}) + P(\text{not satisfied}) - P(\text{new car and not satisfied})$$
$$= \frac{120}{250} + \frac{75}{250} - \frac{28}{250} = \frac{167}{250} \approx 0.668$$

The probability that a person bought a new car or was not satisfied is approximately 0.668 or 66.8%.

Independent Events

Sometimes we need to calculate probabilities for compound events that are connected by the word “and.” Tossing a coin multiple times or rolling dice are independent events. Each time you toss a fair coin the probability of getting heads is ½. It does not matter what happened the last time you tossed the coin. It’s similar for dice. If you rolled double sixes last time that does not change the probability that you will roll double sixes this time. Drawing two cards without replacement is not an independent event. When you draw the first card and set it aside, the probability for the second card is now out of 51 cards not 52 cards.

Definition: Independent Events

Two events are **independent events** if the occurrence of one event has no effect on the probability of the occurrence of the other event.

Example 5.3.6: Determining When Events are Independent

Are these events independent?

a) A fair coin is tossed two times. The two events are (1) first toss is a head and (2) second toss is a head.

b) The two events (1) “It will rain tomorrow in Houston” and (2) “It will rain tomorrow in Galveston” (a city near Houston).

c) You draw a card from a deck, then draw a second card without replacing the first.

Solution

a) The probability that a head comes up on the second toss is $\frac{1}{2}$ regardless of whether or not a head came up on the first toss, so these events are *independent*.

b) These events are *not independent* because it is more likely that it will rain in Galveston on days it rains in Houston than on days it does not.

c) The probability of the second card being red depends on whether the first card is red or not, so these events are *not independent*.

Multiplication Rule for "And" Probabilities: Independent Events

If events *A* and *B* are independent events, then $P(\text{A and B}) = P(A) \cdot P(B)$.

Example 5.3.7: Independent Events for Tossing Coins

Suppose a fair coin is tossed four times. What is the probability that all four tosses land heads up?

Solution

The tosses of the coins are independent events. Knowing a head was tossed on the first trial does not change the probability of tossing a head on the second trial.

$$\begin{aligned} P(\text{four heads in a row}) &= P(\text{1st heads and 2nd heads and 3rd heads and 4th heads}) \\ &= P(\text{1st heads}) \cdot P(\text{2nd heads}) \cdot P(\text{3rd heads}) \cdot P(\text{4th heads}) \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{16} \end{aligned}$$

The probability that all four tosses land heads up is $\frac{1}{16}$.

Example 5.3.8: Independent Events for Drawing Marbles

A bag contains five red and four white marbles. A marble is drawn from the bag, its color recorded and the marble is returned to the bag. A second marble is then drawn. What is the probability that the first marble is red and the second marble is white?

Since the first marble is put back in the bag before the second marble is drawn these are independent events.

$$\begin{aligned} P(\text{1st red and 2nd white}) &= P(\text{1st red}) \cdot P(\text{2nd white}) \\ &= \frac{5}{9} \cdot \frac{4}{9} = \frac{20}{81} \end{aligned}$$

The probability that the first marble is red and the second marble is white is $\frac{20}{81}$.

Example 5.3.9: Independent Events for Faulty Alarm Clocks

Abby has an important meeting in the morning. She sets three battery-powered alarm clocks just to be safe. If each alarm clock has a 0.03 probability of malfunctioning, what is the probability that all three alarm clocks fail at the same time?

Solution

Since the clocks are battery powered we can assume that one failing will have no effect on the operation of the other two clocks. The functioning of the clocks is independent.

$$\begin{aligned} P(\text{all three fail}) &= P(\text{first fails}) \cdot P(\text{second fails}) \cdot P(\text{third fails}) \\ &= (0.03)(0.03)(0.03) \\ &= 2.7 \times 10^{-5} \end{aligned}$$

The probability that all three clocks will fail is approximately 0.000027 or 0.0027%. It is very unlikely that all three alarm clocks will fail.

At Least Once Rule for Independent Events

Many times we need to calculate the probability that an event will happen at least once in many trials. The calculation can get quite complicated if there are more than a couple of trials. Using the complement to calculate the probability can simplify the problem considerably. The following example will help you understand the formula.

Example 5.3.10: At Least Once Rule

The probability that a child forgets her homework on a given day is 0.15. What is the probability that she will forget her homework at least once in the next five days?

Solution

Assume that whether she forgets or not one day has no effect on whether she forgets or not the second day.

If *P*(forgets) = 0.15, then *P*(not forgets) = 0.85.

$$\begin{aligned} P(\text{forgets at least once in 5 tries}) &= P(\text{forgets 1, 2, 3, 4 or 5 times in 5 tries}) \\ &= 1 - P(\text{forgets 0 times in 5 tries}) \\ &= 1 - P(\text{not forget}) \cdot P(\text{not forget}) \cdot P(\text{not forget}) \cdot P(\text{not forget}) \cdot P(\text{not forget}) \\ &= 1 - (0.85)(0.85)(0.85)(0.85)(0.85) \\ &= 1 - (0.85)^5 = 0.556 \end{aligned}$$

The probability that the child will forget her homework at least one day in the next five days is 0.556 or 55.6%

The idea in Example 5.3.9 can be generalized to get the At Least Once Rule.

Definition: At Least Once Rule

If an experiment is repeated n times, the n trials are independent and the probability of event A occurring one time is P(A) then the probability that A occurs at least one time is: $P(\text{A occurs at least once in n trials}) = 1 - P(\overline{A})^n$

Example 5.3.11: At Least Once Rule for Bird Watching

The probability of seeing a falcon near the lake during a day of bird watching is 0.21. What is the probability that a birdwatcher will see a falcon at least once in eight trips to the lake?

Solution

Let A be the event that he sees a falcon so P(A) = 0.21. Then, $P(\overline{A}) = 1 - 0.21 = 0.79$.

$$P(\text{at least once in eight tries}) = 1 - P(\overline{A})^8$$
$$= 1 - (0.79)^8$$
$$= 1 - (0.152) = 0.848$$

The probability of seeing a falcon at least once in eight trips to the lake is approximately 0.848 or 84.8%.

Example 5.3.12: At Least Once Rule for Guessing on Multiple Choice Tests

A multiple choice test consists of six questions. Each question has four choices for answers, only one of which is correct. A student guesses on all six questions. What is the probability that he gets at least one answer correct?

Solution

Let A be the event that the answer to a question is correct. Since each question has four choices and only one correct choice, $P(\text{correct}) = \frac{1}{4}$.

That means $P(\text{not correct}) = 1 - \frac{1}{4} = \frac{3}{4}$.

$$P(\text{at least one correct in six trials}) = 1 - P(\text{not correct})^6$$
$$= 1 - \left(\frac{3}{4}\right)^6$$
$$= 1 - (0.178) = 0.822$$

The probability that he gets at least one answer correct is 0.822 or 82.2%.

Probabilities from Two-Way Tables

Two-way tables can be used to define events and find their probabilities using two different approaches: intuitively or using the probability rules. We can calculate “and” and "or" probabilities by combining the data in relevant cells.

Example 5.3.13: Probabilities from a Two-Way Table

Continuation of Example 5.3.5:

Two hundred fifty people who recently purchased a car were questioned and the results are summarized in the following table.

Table 5.3.1: Satisfaction of Car Buyers

	Satisfied	Not Satisfied	Total
New Car	92	28	120
Used Car	83	47	130
Total	175	75	250

A person is chosen at random. Find the probability that the person:

1. bought a new car

$$P(\text{new car}) = \frac{\text{number of new car}}{\text{number of people}}$$
$$= \frac{120}{250} = 0.480 = 48.0\%$$

2. was satisfied

$$P(\text{satisfied}) = \frac{\text{number of satisfied}}{\text{number of people}}$$
$$= \frac{175}{250} = 0.700 = 70.0\%$$

3. bought a new car and was satisfied

$$P(\text{new car and satisfied}) = \frac{\text{number of new car and satisfied}}{\text{number of people}}$$
$$= \frac{92}{250} = 0.368 = 36.8\%$$

4. bought a new car or was satisfied

$$P(\text{new car or satisfied}) = \frac{\text{number of new car + number of satisfied - number of new car and satisfied}}{\text{number of people}}$$
$$= \frac{120 + 175 - 92}{250} = \frac{203}{250} = 0.812 = 81.2\%$$

This page titled 5.3: Probability Rules- “And” and “Or” is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier via source content that was edited to the style and standards of the LibreTexts platform.

- 3.2: Combining Probabilities with “And” and “Or” by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.
- 3.3: Conditional Probabilities by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.
- 10.2: Working with Events by Darlene Diaz is licensed CC BY-NC-SA 4.0. Original source: https://www.sccollege.edu/OER/Documents/MATH105/Math For Liberal Art Students (2017).pdf.

5.4: Counting Methods

Learning Objectives

Students will be able to:

- Apply the multiplication rule to quickly count the number of possible outcomes.
- Count the number of different possible arrangements by using the permutation rule.
- Count the number of different possible groupings by using the combinations rule.

Recall that

$$P(A) = \frac{\text{number of ways for } A \text{ to occur}}{\text{total number of outcomes}} \tag{5.4.1}$$

for theoretical probabilities. So far the problems we have looked at had rather small total number of outcomes. We could easily count the number of elements in the sample space. If there are a large number of elements in the sample space we can use counting techniques such as permutations or combinations to count them.

Multiplication Principle and Tree Diagrams

The simplest of the counting techniques is the multiplication principle. A tree diagram is a useful tool for visualizing the multiplication principle.

Example 5.4.1: Multiplication Principle for a Three Course Dinner

Let's say that a person walks into a restaurant for a three course dinner. There are four different salads, three different entrees, and two different desserts to choose from. Assuming the person wants to eat a salad, an entrée and a desert, how many different meals are possible?

Solution

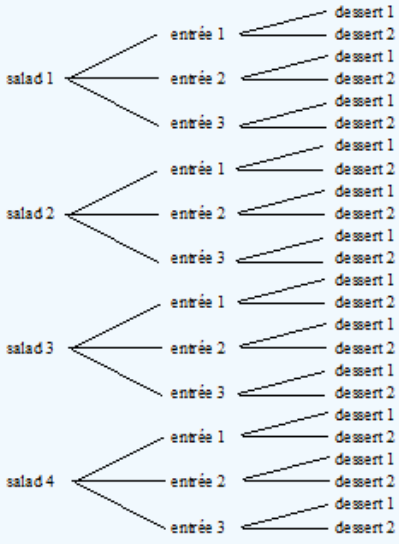


Figure 5.4.1: Tree Diagram for Three-Course Dinner

Looking at the tree diagram we can see that the total number of meals is $4 \times 3 \times 2 = 24$ meals.

Definition: Multiplication Principle

If there are n_1 ways to of choosing the first item, n_2 ways of choosing the second item after the first item is chosen, n_3 ways of choosing the third item after the first two have been chosen, and so on until there are n_k ways of choosing the last item after the earlier choices, then the total number of choices overall is given by

$$n_1 \times n_2 \times n_3 \times n_4 \times n_5 ... \times n_k. \tag{5.4.2}$$

Example 5.4.2: Multiplication Principle for Lining up People

Let's look at the number of ways that four people can line up. We can choose any of the four people to be first. Then there are three people who can be second and two people who can be third. At this point there is only one person left to be last. Using the multiplication principle there are

$$4 \times 3 \times 2 \times 1 = 24 \text{ ways}$$

for four people to line up.

This type of calculation occurs frequently in counting problems so we have some notation to simplify the problem.

Definition: Factorial

The **factorial** of n, read "n factorial" is

$$n! = n(n-1)(n-2)(n-3)...(2)(1). \tag{5.4.3}$$

By this definition, $0! = 1$.

Example 5.4.3: Factorials

$$\begin{aligned} 5! &= 5 \times 4 \times 3 \times 2 \times 1 \\ &= 120 \\ 8! &= 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 \\ &= 40,320 \end{aligned}$$

Factorials get very large very fast.

$$20! = 2.43 \times 10^{18}$$

and

$$40! = 8.16 \times 10^{47}.$$

70! is larger than most calculators can handle.

The multiplication principle may seem like a very simple idea but it is very powerful. Many complex counting problems can be solved using the multiplication principle.

Example 5.4.4: Multiplication Principle for License Plates

Some license plates in Arizona consist of three digits followed by three letters. How many license plates of this type are possible if:

1. There are 10 digits (0, 1, 2, 3, ..., 9) and 26 letters.
$$\underbrace{(10 \cdot 10 \cdot 10)}_{\text{digits}} \cdot (26 \cdot 26 \cdot 26) = 17,576,000 \text{license plates}$$
2. letters can be repeated but digits cannot?
$$\underbrace{(\underline{10} \cdot \underline{9} \cdot \underline{8})}_{\text{letters}} \cdot (26 \cdot 26 \cdot 26) = 12,654,720 \text{ license plates}$$
3. the first digit cannot be zero and both digits and letters can be repeated?
$$\underbrace{(\underline{9} \cdot \underline{10} \cdot \underline{10})}_{\text{digits}} \cdot \underbrace{(\underline{26} \cdot 26 \cdot 26)}_{\text{letters}} = 15,818,400 \text{ license plates}$$
4. neither digits nor numbers can be repeated.
$$\underbrace{(\underline{10} \cdot \underline{9} \cdot \underline{8})}_{\text{digits}} \cdot \underbrace{(\underline{26} \cdot \underline{25} \cdot \underline{24})}_{\text{letters}} = 11,232,000 \text{ license plates}$$

Permutations

Consider the following counting problems:

1. In how many ways can three runners finish a race?
2. In how many ways can a group of three people be chosen to work on a project?

What is the difference between these two problems? In the first problem the order that the runners finish the race matters. In the second problem the order in which the three people are chosen is not important, only which three people are chosen matters.

Definition: Permutation

A **permutation** is an arrangement of a set of items. The number of permutations of n items taking r at a time is given by:

$$P(n,r) = \frac{n!}{(n-r)!} \tag{5.4.4}$$

Note: Many calculators can calculate permutations directly. Look for a function that looks like ${}_nP_r$ or $P(n,r)$

Example 5.4.5: Permutation for Race Cars

Let's look at a simple example to understand the formula for the number of permutations of a set of objects. Assume that 10 cars are in a race. In how many ways can three cars finish in first, second and third place? The order in which the cars finish is important. Use the multiplication principle. There are 10 possible cars to finish first. Once a car has finished first, there are nine cars to finish second. After the second car is finished, any of the eight remaining cars can finish third. 10 x 9 x 8 = 720. This is a permutation of 10 items taking three at a time.

Using the permutation formula (Equation 5.4.4):

$$P(10,3) = \frac{10!}{(10-3)!} = \frac{10!}{7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 10 \cdot 9 \cdot 8 = 720$$

Using the multiplication principle (Equation 5.4.2):

$$\underline{10} \cdot \underline{9} \cdot \underline{8} = 720$$

There are 720 different ways for cars to finish in the top three places.

Example 5.4.6: Permutation for Orchestra Programs

The school orchestra is planning to play six pieces of music at their next concert. How many different programs are possible?

Solution

This is a permutation because they are arranging the songs in order to make the program. Using the permutation formula (Equation 5.4.4):

$$P(6,6) = \frac{6!}{(6-6)!} = \frac{6!}{0!} = \frac{720}{1} = 720$$

Using the multiplication principle (Equation 5.4.2):

$$\underline{6} \cdot \underline{5} \cdot \underline{4} \cdot \underline{3} \cdot \underline{2} \cdot \underline{1} = 720$$

There are 720 different ways of arranging the songs to make the program.

Example 5.4.7: Permutation for Club Officers

The Volunteer Club has 18 members. An election is held to choose a president, vice-president and secretary. In how many ways can the three officers be chosen?

Solution

The order in which the officers are chosen matters so this is a permutation.

Using the permutation formula (Equation 5.4.4):

$$P(18,3) = \frac{18!}{(18-3)!} = \frac{18!}{15!} = 18 \cdot 7 \cdot 6 = 4896$$

Note: All digits in 18! in the numerator from 15 down to one will cancel with the 15! in the denominator.

Using the multiplication principle (Equation 5.4.2):

$$\underset{\text{Pres.}}{\underline{18}} \cdot \underset{\text{V.P.}}{\underline{17}} \cdot \underset{\text{Sec.}}{\underline{16}} = 4896$$

There are 4896 different ways the three officers can be chosen.

Another notation for permutations is ${}_nP_r$. So, $P(18,3)$ can also be written as ${}_{18}P_3$. Most scientific calculators have an ${}_nP_r$ button or function.

Combinations

Example 5.4.8: Formula for Combinations

Choose a committee of two people from persons A, B, C, D and E. By the multiplication principle there are *5 \cdot 4 = 20\) ways to arrange the two people.

AB AC AD AE BA BC BD BE CA CB

CD CE DA DB DC DE EA EB EC ED

Committees AB and BA are the same committee. Similarly for committees CD and DC. Every committee is counted twice.

$$\frac{20}{2} = 10$$

so there are 10 possible different committees.

Now choose a committee of three people from persons A, B, C, D and E. There are $5 \cdot 4 \cdot 3 = 60$ ways to pick three people in order. Think about the committees with persons A, B and C. There are $3! = 6$ of them.

ABC ACB BAC BCA CAB CBA

Each of these is counted as one of the 60 possibilities but they are the same committee. Each committee is counted six times so there are

$$\frac{60}{6} = 10 \text{ different committees.}$$

In both cases we divided the number of permutations by the number of ways to rearrange the people chosen.

The number of permutations of n people taking r at a time is $P(n,r)$ and the number of ways to rearrange the people chosen is $r!$. Putting these together we get

$$\begin{aligned} \frac{n!}{\#\text{ ways to arrange r items}} &= \frac{P(n,r)}{r!} = \frac{(n-r)!}{\frac{r!}{1}} \\ &= \frac{n!}{(n-r)!} \cdot \frac{1}{r!} \\ &= \frac{n!}{(n-r)!r!} \end{aligned} \tag{5.4.5}$$

Definition: Combination

A **combination** is a selection of objects in which the order of selection does not matter. The number of combinations of n items taking r at a time is:

$$C(n,r) = \frac{n!}{r!(n-r)!} \tag{5.4.6}$$

Note: Many calculators can calculate combinations directly. Look for a function that looks like ${}_nC_r$ or (C(n,r)).

Example 5.4.9: Combination for Picking Books

A student has a summer reading list of eight books. The student must read five of the books before the end of the summer. In how many ways can the student read five of the eight books?

Solution

The order of the books is not important, only which books are read. This is a combination of eight items taking five at a time.

$$C(8,5) = \frac{8!}{5!(8-5)!} = \frac{8!}{5!3!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 8 \times 7 = 56 \tag{5.4.7}$$

There are 56 ways to choose five of the books to read.

Example 5.4.10: Combination for Halloween Candy

A child wants to pick three pieces of Halloween candy to take in her school lunch box. If there are 13 pieces of candy to choose from, how many ways can she pick the three pieces?

Solution

This is a combination because it does not matter in what order the candy is chosen.

$$\begin{aligned} C(13,3) &= \frac{13!}{3!(13-3)!} = \frac{13!}{3!10!} \\ &= \frac{13\times 12\times 11\times 10\times 9\times 8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}{(3\times 2\times 1)(10\times 9\times 8\times 7\times 6\times 5\times 4\times 3\times 2\times 1} \\ &= \frac{13\times 12\times 11}{3\times 2\times 1} \\ &= \frac{1716}{6} = 286 \end{aligned}$$

There are 286 ways to choose the three pieces of candy to pack in her lunch.

Note: The difference between a combination and a permutation is whether order matters or not. If the order of the items is important, use a permutation. If the order of the items is not important, use a combination.

Example 5.4.11: Permutation or Combination for Bicycle Serial Numbers

A serial number for a particular model of bicycle consists of a letter followed by four digits and ends with two letters. Neither letters nor numbers can be repeated. How many different serial numbers are possible?

Solution

This is a permutation because the order matters.

Use the multiplication principle to solve this. There are 26 letters and 10 digits possible.

$$26\cdot 10\cdot \underline{9}\cdot \underline{8}\cdot \underline{7}\cdot \underline{25}\cdot \underline{24} = 78,624,000 \tag{5.4.8}$$

There are 78,624,000 different serial numbers of this form.

Example 5.4.12: Permutation or Combination for Choosing Men and Women

A class consists of 15 men and 12 women. In how many ways can two men and two women be chosen to participate in an in-class activity?

Solution

This is a combination since the order in which the people is chosen is not important.

Choose two men:

$$C(15,2) = \frac{15!}{2!(15-2)!} = \frac{15!}{2!13!} = 105$$

Choose two women:

$$C(12,2)=\frac{12!}{2!(12-2)!}=\frac{12!}{2!10!}=66$$

We want 2 men and 2 women so multiply these results.

$$105(66)=6930$$

There are 6930 ways to choose two men and two women to participate.

Probabilities Involving Permutations and Combinations

Now that we can calculate the number of permutations or combinations, we can use that information to calculate probabilities.

Example 5.4.13: Probability with a Combination for Choosing Students

There are 20 students in a class. Twelve of the students are women. The names of the students are put into a hat and five names are drawn. What is the probability that all of the chosen students are women?

Solution

This is a combination because the order of choosing the students is not important.

$$P(\text{ all females })=\frac{\#\text{ ways to pick 5 women}}{\#\text{ ways to pick 5 students}} \tag{5.4.9}$$

The number of way to choose 5 women is

$$C(12,5)=792$$

The number of ways to choose 5 students is

$$C(20,5)=15,504$$

$$P(\text{ all females })=\frac{\#\text{ ways to pick 5 women}}{\#\text{ ways to pick 5 students}}=\frac{792}{15,504}=0.051$$

The probability that all the chosen students are women is 0.051 or 5.1%.

Example 5.4.14: Probability with a Permutation for a Duck Race

The local Boys and Girls Club holds a duck race to raise money. Community members buy a rubber duck marked with a numeral between 1 and 30, inclusive. The box of 30 ducks is emptied into a creek and allowed to float downstream to the finish line. What is the probability that ducks numbered 5, 18 and 21 finish in first, second and third, respectively?

Solution

This is a permutation since the order of finish is important.

$$P(5,18\&21\text{ finish 1 st, 2 nd \& 3 rd})=\frac{\#\text{ ways }5,18\&21\text{ finish 1 st, 2 nd \&3rd}}{\#\text{ ways any ducks can finish 1 st, 2 nd \& 3rd}}$$

There is only one way that the ducks can finish with #5 in first, #18 in second and #21 in third.

The number of ways any ducks can finish in first, second and third is

$$P(30,3)=24,360$$

$$P(5, 18\&21 \text{ finish 1 st, 2 nd \&3rd}) = \frac{\#\text{ ways } 5, 18\&21 \text{ finish 1 st, 2 nd \&3 rd}}{\#\text{ ways any ducks can finish 1st,, 2nd\&3 rd}}$$
$$= \frac{1}{24,360} \approx 4.10x10^{-5}$$

The probability that ducks numbered 5, 18 and 21 finish in first, second and third, respectively, is approximately 0.000041 or 0.0041%.

Example 5.4.15: Probability with a Permutation for Two-Card Poker Hands

A poker hand consists of two cards. What is the probability that the poker hand consists of two jacks or two fives?

Solution

It's not possible to get two jacks and two fives at the same time so these are mutually exclusive events.

The number of ways to get two jacks is

$$C(4, 2) = 6. \tag{5.4.10}$$

The number of ways to get two fives is

$$C(4, 2) = 6 \tag{5.4.11}$$

The number of ways to get two jacks or two fives is

$$6 + 6 = 12 \tag{5.4.12}$$

The total number of ways to get a 2-card poker hand is

$$C(52, 2) = 1326 \tag{5.4.13}$$

$$P(2 \text{ jacks or } 2 \text{ fives }) = \frac{\text{number of ways to get 2 jacks or 2 fives}}{\text{number of ways to choose 2 cards}} = \frac{12}{1326} \approx 0.009 \tag{5.4.14}$$

The probability of getting two jacks or two fives is approximately 0.009 or 0.9%.

Example 5.4.16: Probability with a Combination for Rotten Apples

A basket contains 10 good apples and two bad apples. If a distracted shopper reaches into the basket and picks three apples without looking, what is the probability he gets one bad apple?

Solution

This is a combination since the order in which the apples were picked is not important. He picks three apples total. If one apple is bad the other two must be good. Find the probability of one bad apple and two good apples.

$$P(\text{ one bad and two good apples }) = \frac{\#\text{ ways to get one bad and two good apples}}{\#\text{ ways to get three apples}}$$

The number of ways to get one bad and two good apples is

$$C(2, 1) \cdot C(10, 2) = 2 \cdot 45 = 90$$

The number of ways to get three apples is

$$C(10, 3) = 120 \tag{5.4.15}$$

$$P(\text{ one bad and two good apples }) = \frac{\#\text{ ways to get one bad and two good apples}}{\#\text{ ways to get three apples}}$$
$$= \frac{90}{120}$$
$$= 0.75$$

The probability of getting one bad apple out of three apples is 0.75 or 75%

This page titled 5.4: Counting Methods is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier via source content that was edited to the style and standards of the LibreTexts platform.

- 3.5: Counting Methods by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.

5.5: Conditional Probabilities

Learning Objectives

Students will be able to:

- Calculate the conditional probability of an event A given that event B has already occurred.
- Identify instances when the occurance of one event will influence the probably of another event.
- Compute the conditional probability of an event by using a two-way table.

What do you think the probability is that a man is over six feet tall? If you knew that both his parents were tall would you change your estimate of the probability? A conditional probability is a probability that is based on some prior knowledge.

Definition: Conditional Probability

A **conditional probability** is the probability that an event will occur if some other condition has already occurred. This is denoted by $P(A|B)$, which is read "the probability of *A* given *B*."

Example 5.5.1 Conditional Probability for Drawing Cards without Replacement

Two cards are drawn from a well shuffled deck of 52 cards without replacement. Find the following probabilities:

a. The probability that the second card is a heart given that the first card is a spade.
b. The probability that the first card is a face card and the second card an ace.
c. The probability that one card is a heart and the other a club.

Solution a

Without replacement means that the first card is set aside before the second card is drawn and we assume the first card is a spade. There are only 51 cards to choose from for the second card. Thirteen of those cards are hearts.

It's important to notice that the question only asks about the second card.

$$P(\text{2nd heart} \mid \text{1st spade}) = \frac{13}{51}$$

The probability that the second card is a heart given that the first card is a spade is $\frac{13}{51}$.

Solution b

Notice that this time the question asks about both of the cards.

There are 12 face cards out of 52 cards when we draw the first card. We set the first card aside and assume that it is a face card. Then there are four aces out of the 51 remaining cards. We want to draw a face card and an ace so use multiplication.

$$P(\text{1st face card and 2nd ace}) = \frac{12}{52} \cdot \frac{4}{51} = \frac{48}{2652} \approx 0.018$$

The probability that the first card is a face card and the second card an ace is approximately 0.018 or 1.8%.

Solution c

There are two ways for this to happen. We could get a heart first and a club second or we could get the club first and the heart second.

$$\begin{aligned} P(\text{heart and club}) &= P(\text{heart 1st and club 2nd or club 1st and heart 2nd}) \\ &= P(\text{heart 1st and club 2nd}) + P(\text{club 1st and heart 2nd}) \\ &= \frac{13}{52} \cdot \frac{13}{51} + \frac{13}{52} \cdot \frac{13}{51} \\ &\approx 0.127 \end{aligned}$$

The probability that one card is a heart and the other a club is approximately 0.127 or 12.7%.

Example 5.5.2 Conditional Probability for Rolling Dice

Two fair dice are rolled and the sum of the numbers is observed. What is the probability that the sum is at least nine if it is known that one of the dice shows a five?

Since we are given that one of the dice shows a five this is a conditional probability. List the pairs of dice with one die showing a five. Be careful not to count (5,5) twice.

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

List the pairs from above that have a sum of at least nine.

$$\{(4,5),(5,5),(6,5),(5,4),(5,6)\}$$

Solution

There are 11 ways for one die to show a five and five of these ways have a sum of at least nine.

$$P(\text{sum at least 9} \mid \text{one die is a 5}) = \frac{5}{11}$$

The probability that the sum is at least nine if it is known that one of the dice shows a five is $\frac{5}{11}$.

Multiplication Rule for "And" Probabilities: Any Events

For events A and B, $P(A \text{ and } B) = P(A) \cdot P(B|A)$

Conditional Probability

For events A and B, $P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$

Example 5.5.3: Conditional Probability for Satisfaction of Car Buyers

Two hundred fifty people who recently purchased a car were questioned and the results are summarized in the following table from Section 5.3.

Table 5.5.1: Satisfaction of Car Buyers

	Satisfied	Not Satisfied	Total
New Car	92	28	120

	Satisfied	Not Satisfied	Total
Used Car	83	47	130
Total	175	75	250

This is a conditional probability because we already know that the person bought a used car.What is the probability that a person is satisfied if it is known that the person bought a used car?

Solution

$$P(\text{satisfied} \mid \text{used car}) = \frac{P(\text{satisfied and used})}{P(\text{used})}$$
$$= \frac{\frac{83}{250}}{\frac{130}{250}} = \frac{83}{\cancel{250}} \cdot \frac{\cancel{250}}{130} = \frac{83}{130} \approx 0.638$$

The probability that a person is satisfied if it is known that the person bought a used car is approximately 0.638 or 63.8%

Example 5.5.4: Conditional Probability for Residence and Class Standing

A survey of 350 students at a university revealed the following data about class standing and place of residence.

Table 5.5.2: Housing by Class

Residence\Class	Freshman	Sophomore	Junior	Senior	Row Totals
Dormitory	89	34	46	15	184
Apartment	32	17	22	48	119
With Parents	13	31	3	0	47
Column Totals	134	82	71	63	350

What is the probability that a student is a sophomore if that student lives in an apartment?

Solution

This is a conditional probability because we are given that the student lives in an apartment.

$$P(\text{sophomore} \mid \text{apartment}) = \frac{P(\text{sophomore and apartment})}{P(\text{apartment})}$$
$$= \frac{\frac{17}{350}}{\frac{119}{350}} = \frac{17}{\cancel{350}} \cdot \frac{\cancel{350}}{119} = \frac{17}{119} \approx 0.143$$

The probability that a student is a sophomore given that the student lives in an apartment is approximately 0.143 or 14.3%.

This page titled 5.5: Conditional Probabilities is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier via source content that was edited to the style and standards of the LibreTexts platform.

- 3.3: Conditional Probabilities by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.

6: INFERENTIAL STATISTICS



- Topic hierarchy

6.1: Discrete Random Variables

6.2: Continuous Random Variables

6.3: Normal Distributions

6.4: Estimating Population Mean

6.5: Estimating Population Proportion

This page titled 6: Inferential Statistics is shared under a license and was authored, remixed, and/or curated by LibreTexts.

6.1: Discrete Random Variables

SECTION OBJECTIVES

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

- Recognize and use discrete random variables.
- Calculate and interpret expected values.

- A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?
- Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

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



Figure 6.1.1 You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. If X is a random variable and x is given as a number then $X = x$ is an event with its probability expressed as $P(X = x)$.

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

Discrete Random Variables

A discrete probability distribution table (function) lists possible values of the random variable along with the probability of each outcome. It has two characteristics:

a. Each probability is between zero and one, inclusive.
b. The sum of the probabilities is one.

Example 6.1.1

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

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

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

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

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

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1 \tag{6.1.1}$$

Exercise 6.1.1

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

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

Answer

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

$$\frac{4}{50} + \frac{8}{50} + \frac{16}{50} + \frac{14}{50} + \frac{6}{50} + \frac{2}{50} = 1 \tag{6.1.2}$$

Interactive Exercise 6.1.1

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

View this Interactive Content online

Example 6.1.2

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

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

Solutions

a. Let X be the number of days Nancy attends class per week.

b. 0, 1, 2, and 3

c

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

Exercise 6.1.2

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

Answer

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

Interactive Exercise 6.1.2

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

View this Interactive Content online

Example 6.1.3

A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution. Let X be the number of years a new hire will stay with the company. Let $P(X = x)$ be the probability that a new hire will stay with the company x years.

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

1. Complete the table using the data provided, i.e. find $P(X = 4)$.

2. Find $P(X \geq 5)$.

Answer

1. $P(X = 4) = 1 - (0.12 + 0.18 + 0.30 + 0.15 + 0.10 + 0.05) = 1 - 0.90 = 0.10$ thus the complete probability distribution table for X:

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

2. $P(X \geq 5) = 0.10 + 0.05 = 0.15$

Exercise 6.1.3

Javier volunteers in community events each month. He does not do more than five events in a month. Over the years, he has established the following probability distribution. Let X be the number of times Javier volunteers in a randomly selected month and let $P(X = x)$ be the probability that Javier volunteers x times.

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

1. Complete the table using the data provided, i.e. find $P(X = 2)$.

2. Find $P(X \leq 3)$.

Answer available at the end of this section.

Interactive Exercise 6.1.3

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

View this Interactive Content online

Expected Value

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

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

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

To find the **expected value** or long-term average, μ, simply multiply each value of the random variable by its probability and add the products. It is much easier to learn by seen the process rather than looking at the formula.

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. The expected value, μ, in this case is the long-term average of the number of days per week the men's soccer team plays soccer. To find it we create the probability distribution table and add a column $x * P(x)$. In this column, you will multiply each x value by its probability.

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

x	$P(X = x)$	$x \cdot P(X = x)$
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Add the last column $x \cdot P(X = x)$ to find the long term average or expected value:

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

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

Example 6.1.4

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight.

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

x	$P(X = x)$	$x \cdot P(X = x)$
0	$P(X = 0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$
1	$P(X = 1) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$
2	$P(X = 2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$
3	$P(X = 3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$
4	$P(X = 4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$
5	$P(X = 5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$

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

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

Exercise 6.1.4

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

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

Answer available at the end of this section.

Interactive Exercise 6.1.4

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

View this Interactive Content online

Example 6.1.5

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

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

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

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

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

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

$$1 - 0.00001 = 0.99999.1 - 0.00001 = 0.99999.$$

The expected value table is as follows:

Add the last column. –1.99998 + 1 = –0.99998

	x	$P(X = x)$	$x \cdot P(X = x)$
Loss	–2	0.99999	(–2)(0.99999) = –1.99998
Profit	100,000	0.00001	(100000)(0.00001) = 1

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

Exercise 6.1.5

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

a. Define a random variable X.

b. Complete the following expected value table.

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

c. What is the expected value, μ? Do you come out ahead?

Answer available at the end of this section.

Interactive Exercise 6.1.5

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

View this Interactive Content online

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

Answers

Exercise 6.1.3

1. $P(X = 2) = 1 - (0.05 + 0.05 + 0.20 + 0.25 + 0.35) = 1 - 0.90 = 0.10$ thus the complete probability distribution table for X:

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

2. $P(X \leq 3) = 0.05 + 0.05 + 0.10 + 0.20 = 0.40$.

Exercise 6.1.4

The expected value is 2.24

$$(0)\frac{4}{50} + (1)\frac{8}{50} + (2)\frac{16}{50} + (3)\frac{14}{50} + (4)\frac{6}{50} + (5)\frac{2}{50} = 0 + \frac{8}{50} + \frac{32}{50} + \frac{42}{50} + \frac{24}{50} + \frac{10}{50} = \frac{116}{50} = 2.32 \tag{6.1.3}$$

Exercise 6.1.5

a. X = amount of profit

b.

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

c. Add the last column of the table. The expected value $\mu = -\frac{2}{3} = -0.67$. You lose, on average, about 67 cents each time you play the game, so you do not come out ahead.

This page titled 6.1: Discrete Random Variables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- CustomerActivityDetailed has no license indicated.

6.2: Continuous Random Variables

SECTION OBJECTIVES

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

- Recognize and use continuous random variables.
- Compute and interpret probabilities for a continuous random variable.

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long-distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.



Figure 6.2.1: The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Properties of Continuous Probability Distributions

The fact that it is impossible to list all values of a continuous random variable makes it impossible to construct a probability distribution table, so instead, we are going to focus on its visual representation called a **probability density function** (pdf) whose graph is always on or above the horizontal axis and the total area between the graph and the horizontal axis is equal to 1. We say that a continuous random variable X is associated with a given probability density curve $f(x)$ if the probability that X is between a and b, i.e. $P(a < X < b)$, is equal to the area under the $f(x)$ between a and b for any a and b.

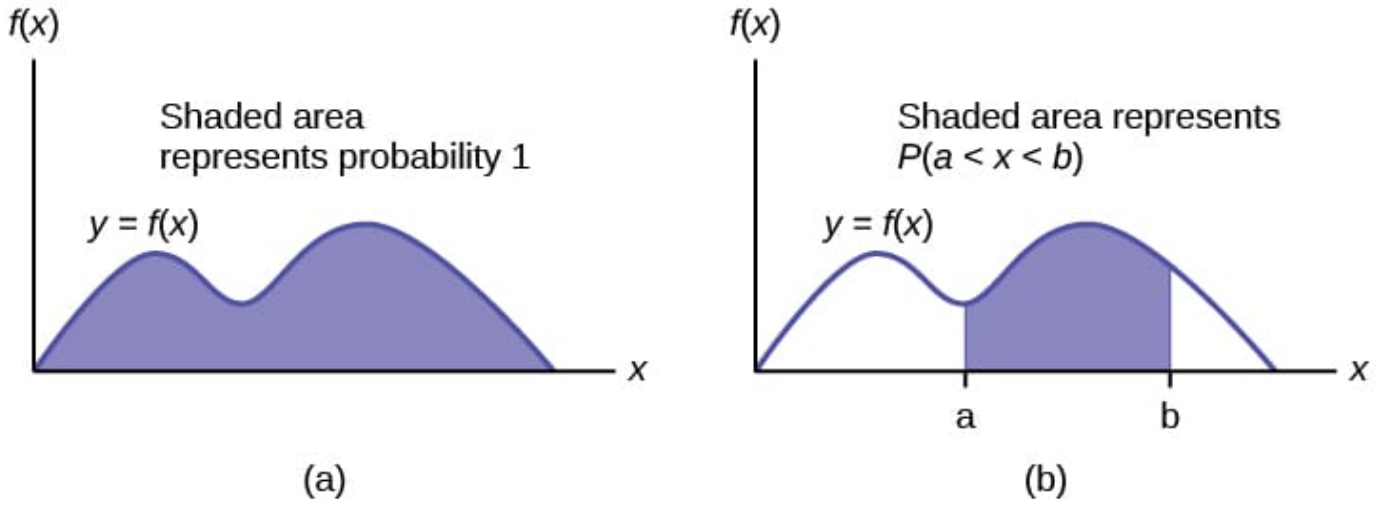


Figure 6.2.2: An example of a probability density function.

In summary:

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- The probability that the random variable X is in the interval between the values c and d, $P(c < X < d)$, is the area under the curve, above the *x*-axis, to the right of c and the left of d.
- The probability that x takes on any single individual value is zero, i.e. $P(X = c) = 0$ for any c. The reason is that the area between the curve and the *x*-axis between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < X < d)$ is the same as $P(c \leq X \leq d)$ because probability is equal to area.
- The complementary rule holds i.e. $P(X < c) + P(X > c) = 1$.

Example 6.2.1

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$.

Exercise 6.2.1

If the area to the left of x is 0.012, then what is the area to the right?

Answer available at the end of this section.

Interactive Exercise 6.2.1

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

View this Interactive Content online

The area that represents probability can be found by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook. There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

The following graphs illustrate the probability density functions and their applications for some common distributions.

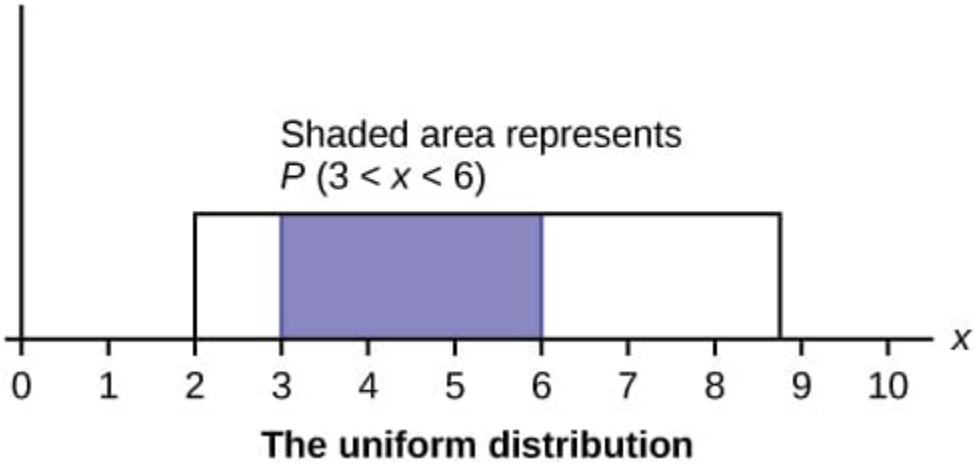


Figure 6.2.3: The graph shows a Uniform Distribution with the area between $x = 3$ and $x = 6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.

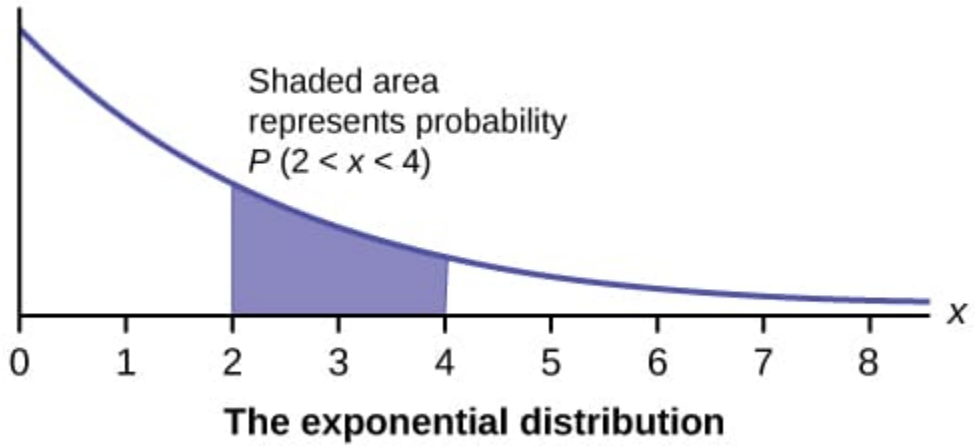


Figure 6.2.4: The graph shows an Exponential Distribution with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.

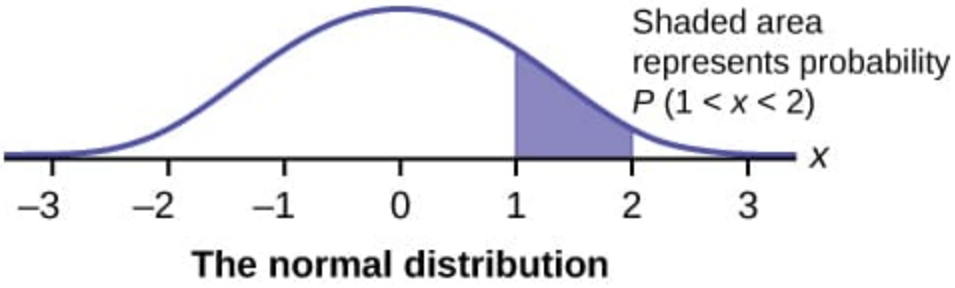


Figure 6.2.5: The graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

Next, we will consider one of the standard continuous random variables called uniform and some of its applications. A **uniform probability density function** with parameters a and b is $f(x) = \frac{1}{b-a}$ whose graph is a horizontal line $f(x) = \frac{1}{b-a}$ above the x-axis from $x = a$ to $x = b$. A random variable X with such a probability density function is called a **uniform random variable** with parameters a and b and we denote it as $X \sim U(a, b)$. The parameters a and b are frequently called the minimum and maximum bounds.

Since the graph of a uniform probability density function is a straight horizontal line, the probability $P(c < X < d)$ is the area of the rectangular region and thus its area can always be found using the formula:

$P(c < X < d) = \frac{d-c}{b-a}$

Example 6.2.2

Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.



Figure 6.2.1

$f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$.

The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the *x*-axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.

$$AREA = 20\left(\frac{1}{20}\right) = 1$$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the *x*-axis where $0 < x < 2$.



Figure 6.2.2

$$AREA = (2-0)\left(\frac{1}{20}\right) = 0.1$$

$$(2-0) = 2 = \text{base of a rectangle}$$

REMINDER: area of a rectangle = (base)(height).

The area corresponds to a probability. If X has the probability density function $f(x)$ then the probability that X is between zero and two is 0.1, which can be written mathematically as $P(0 < X < 2) = P(X < 2) = 0.1$.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the *x*-axis where $4 < x < 15$.

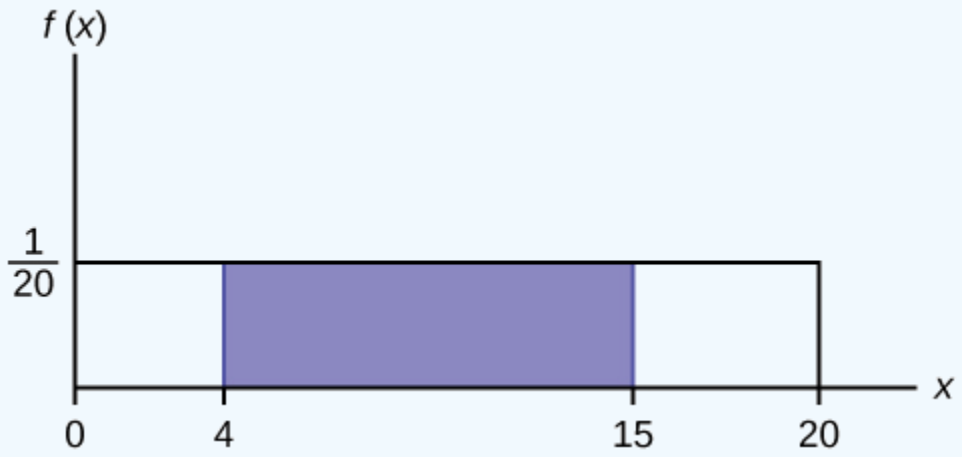


Figure 6.2.3

$\text{AREA} = (15–4)(\frac{1}{20}) = 0.55$

$\text{AREA} = (15–4)(\frac{1}{20}) = 0.55$

$(15–4) = 11 = \text{the base of a rectangle}(15–4) = 11 = \text{the base of a rectangle}$

The area corresponds to the probability $P(4 < X < 15) = 0.55$.

Suppose we want to find $P(X = 15)$. A vertical line $x = 15$ has no width (or zero width). Therefore, $P(X = 15) = (\text{base})(\text{height}) = (0)\left(\frac{1}{20}\right) = 0$

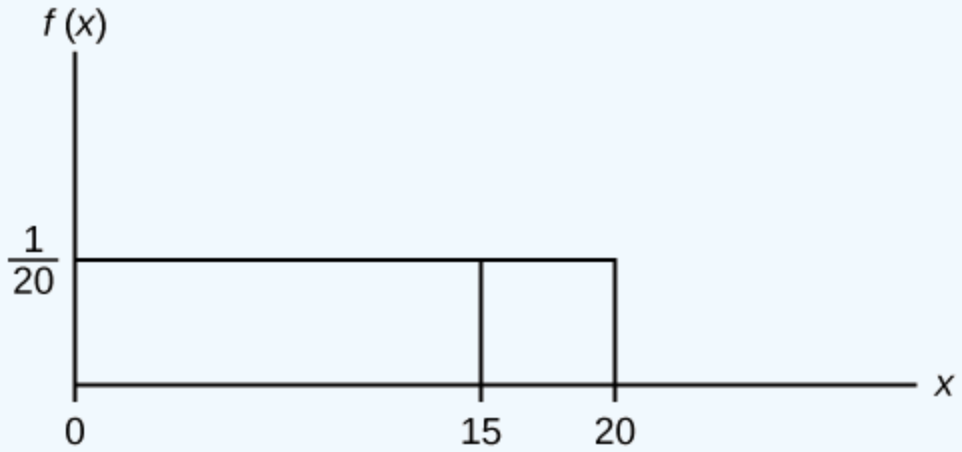


Figure 6.2.4

To calculate the probability that x is between two values, look at the following graph. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.

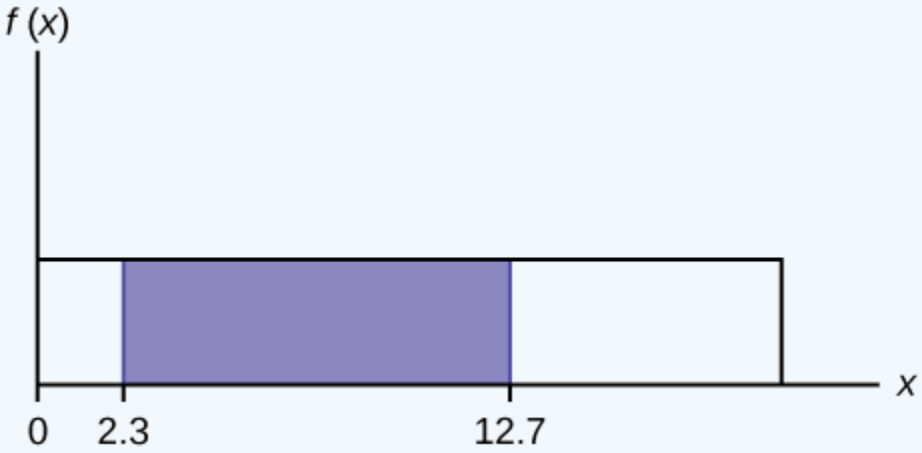


Figure 6.2.5

$P(2.3 < X < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52$

Exercise 6.2.2

Consider the function $f(x) = \frac{1}{8}$ for $0 \leq x \leq 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

Answer available at the end of this section.

Interactive Exercise 6.2.2

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

View this Interactive Content online

Uniform random variables have many applications.

Example 6.2.3

A random passenger comes to a train station through which his desired city metro train runs every 15 minutes. Find the probability that the waiting time is

(i) less than 10 minutes.

(ii) between 3 and 7 minutes.

(iii) more than 12 minutes.

Solution

The waiting time of the passenger can be estimated by a uniform random variable. Let X be the waiting time of a randomly selected passenger then it can be assumed uniform with parameters 0 and 15, i.e. $X \sim U(0, 15)$. Thus:

(i) the probability that the waiting time is less than 10 minutes:

$$P(X < 10) = \text{width} \cdot \text{height} = (10 - 0) \cdot \tfrac{1}{15-0} = \tfrac{10}{15} = 0.67$$

(ii) the probability that the waiting time is between 3 and 7 minutes:

$$P(3 < X < 7) = \text{width} \cdot \text{height} = (7 - 3) \cdot \tfrac{1}{15-0} = \tfrac{4}{15} = 0.27$$

(iii) the probability that the waiting time is more than 12 minutes:

$$P(X > 12) = \text{width} \cdot \text{height} = (15 - 12) \cdot \tfrac{1}{15-0} = \tfrac{3}{15} = 0.20$$

Exercise 6.2.3

A city metro train on the yellow line runs every 10 minutes during the time between 12 pm and 3 pm. Let XX be the waiting time of a randomly selected passenger between 12 pm and 3 pm. Use this information to find the probability that it takes:

i. less than 6 minutes to wait for the train?

ii. more than 7 minutes to wait for the train?

iii. between 4 and 9 minutes to wait for the train?

Answer available at the end of this section.

Interactive Exercise 6.2.3

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

View this Interactive Content online

Answers

Exercise 6.2.1

$1 - 0.012 = 0.988$

Exercise 6.2.2

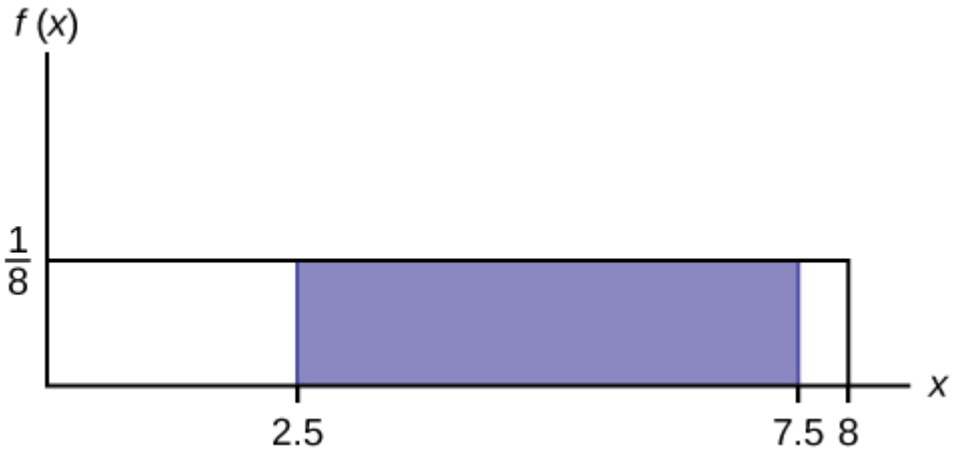


Figure 6.2.6

$P(2.5 < X < 7.5) = 0.625$

Exercise 6.2.3

(i) 0.6 (ii) 0.3 (iii) 0.5

This page titled 6.2: Continuous Random Variables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

6.3: Normal Distributions

SECTIONS OBJECTIVES

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

- Recognize and use the normal probability distribution.
- Recognize and use the standard normal probability distribution.

A special continuous distribution, called normal, is the most common and therefore most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.



Figure 6.3.1: If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

In this section, you will study the normal distribution, the standard normal distribution, and applications associated with them. To study normal distributions, we must introduce a normal probability density function and its graph. Normal distributions have two parameters, the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we denote it as $X \sim N(\mu, \sigma)$ and the probability density function of X is

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{1}{2}\right) \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

As you can see, the probability density function is a rather complicated function. **Do not memorize it** as it is not necessary. So instead, let's focus on its graph which varies depending on the values of μ and σ.

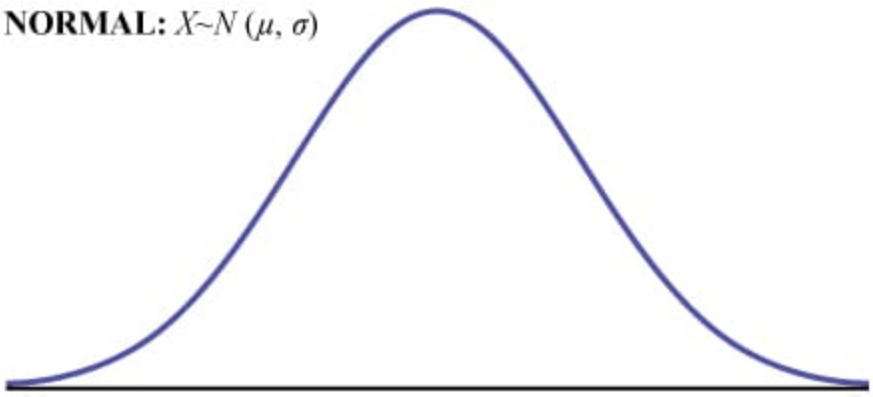


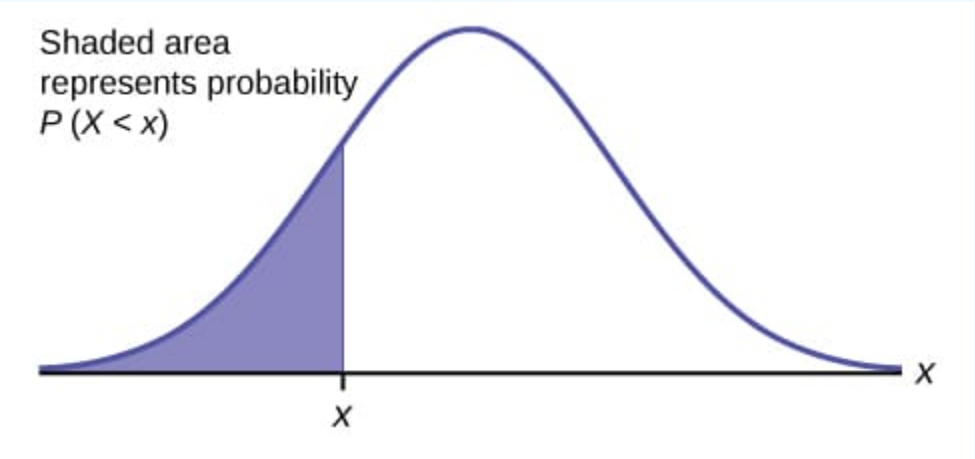
Figure 6.3.2: Normal distribution

Note that the two parameters uniquely determine the shape of the probability density curve, i.e. the center and the spread. The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on the change of σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution,** $Z \sim N(0, 1)$.

Since normal random variables are continuous random variables, everything that we learned about continuous random variables in the previous section will hold for normal variables as well. For example, one of the most important facts is that the area under the probability density curve is the probability of some event that can be expressed in teh inequality form.

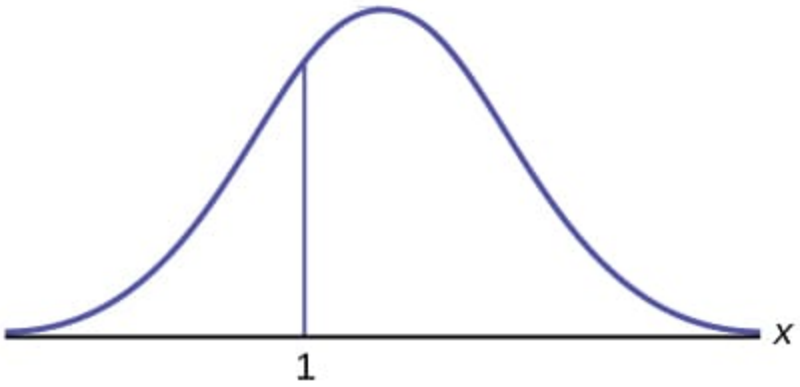
Example 6.3.1

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$.



Exercise 6.3.1.1

How would you represent the area to the left of one in a probability statement?



Answer

$P(X < 1)$

Interactive Exercise 6.3.1.1

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

View this Interactive Content online

Since the $P(X = x) = 0$ for any continuous random variable X and any value x, we conclude that $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for normal distributions as well.

Exercise 6.3.1.2

Is $P(X < 1)$ equal to $P(X \leq 1)$? Why?

Answer

Yes, because they are the same in a continuous distribution: $P(X = 1) = 0$

Interactive Exercise 6.3.1.2

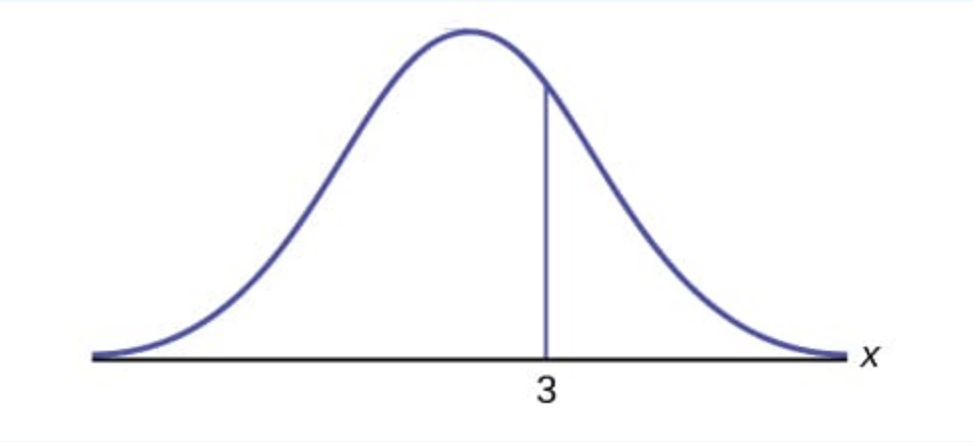
This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Since the total area under the probability density curve is 1, the area to the right is then $P(X > x) = 1 – P(X < x)$. Remember, $P(X < x) =$ **Area to the left** of the vertical line through x. $P(X > x) = 1 – P(X < x) =$ **Area to the right** of the vertical line through x.

Exercise 6.3.1.3

Express symbolically the area to the right of three:



Answer

$1 – P(X < 3)$ or $P(X > 3)$

Interactive Exercise 6.3.1.3

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

View this Interactive Content online

Exercise 6.3.1.4

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Answer

$1 - 0.543 = 0.457$

Interactive Exercise 6.3.1.4

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

View this Interactive Content online

Calculations of Probabilities

Normal tables, computers, and calculators are used to provide or calculate the probabilities such as $P(a < X < b)$, $P(X < c)$, or $P(X > c)$ for any variable X and any values a, b, c. Technology has made the tables virtually obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions in this textbook.

Which Calculator to Use?

There are many online calculators that can be used to compute the probabilities that involve normal and other random variables. For example, you can use this one:

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

View this Interactive Content online

Also, you are encouraged to ask your instructor about which calculator is allowed/recommended for this course.

Example 6.3.2

Use the calculator provided above to verify the following probability statements for $X \sim N(6, 2)$:

$$P(X < 3.5) = 0.1056$$

$$P(4.1 < X < 6.3) = 0.3885$$

$$P(X > 8.2) = 0.1357$$

Exercise 6.3.2

$X \sim N(-3, 4)$

Find the probability that X is between one and four, i.e. $P(1 < X < 4)$.

Answer

0.1186

Interactive Exercise 6.3.2.1

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

View this Interactive Content online

Interactive Exercise 6.3.2.2

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

View this Interactive Content online

Interactive Exercise 6.3.2.3

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

View this Interactive Content online

Recall that the k-th percentile of a dataset is the value such that k% of the data are below this value. In the context of a random variable, the k-th percentile of X is the value x such that the probability $P(X < x) = k/100$.

Example 6.3.3

For $X \sim N(35, 5)$, $P(X < 36) = 0.58$. Therefore 36 is the 58-th percentile of X. If we were asked to find the 58-th percentile of X then essentially we are asked to find the value x such that $P(X < x) = 0.58$. Many online calculators allow to set up such an equation by leaving blank instead of x.

Exercise 6.3.3

Find the 42-nd percentile for a normal random variable $X \sim N(\mu = 93.8, \sigma = 5.8)$.

Answer available at the end of this section.

Interactive Exercise 6.3.3

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

View this Interactive Content online

As we indicated above, normal distributions have miryad of applications.

Example 6.3.4

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.

c. Find the 90th percentile (that is, find the score k that has 90% of the scores below *k* and 10% of the scores above k).
d. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Answer

a. Let X be a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$. Draw a graph. Then, find $P(x > 65)$.

$$P(X > 65) = 0.3446$$

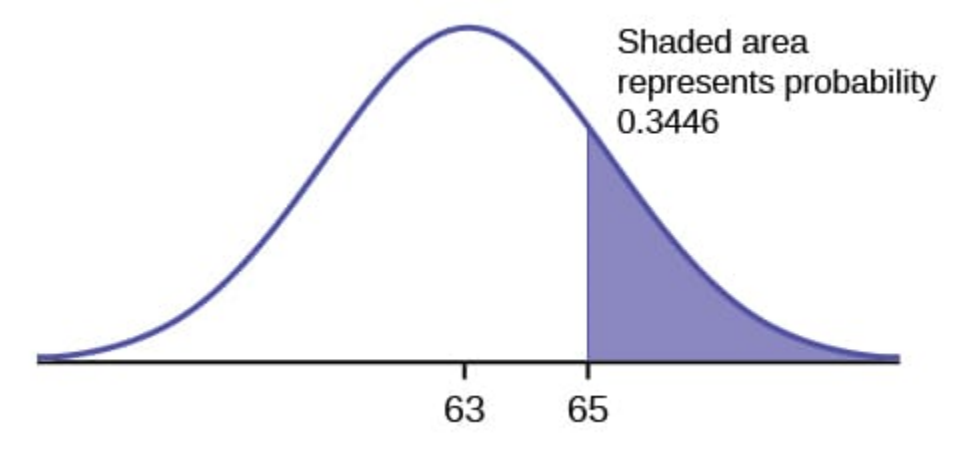


Figure 6.3.3.

The probability that any student selected at random scores more than 65 is 0.3446.

Answer

b. Draw a graph. Then find $P(X < 85)$, and shade the graph. Using a computer or calculator, find $P(X < 85) = 1$. The probability that one student scores less than 85 is approximately one (or 100%).

Answer

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let k be the 90th percentile. The variable k is located on the x-axis. $P(X < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher. The variable k is often called a critical value. Using a computer or calculator, $k = 69.4$

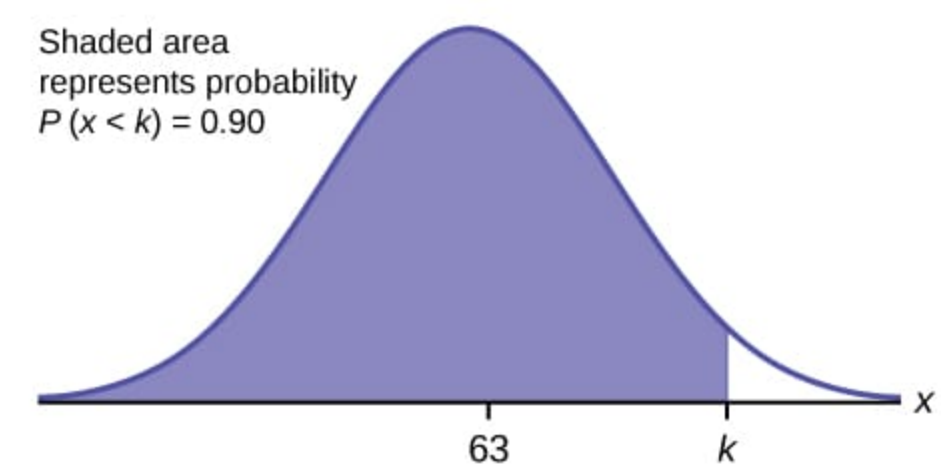


Figure 6.3.4.

The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above.

Answer

d. Find the 70th percentile.

Draw a new graph and label it appropriately. The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

Exercise 6.3.4

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Answer available at the end of this section.

Interactive Exercise 6.3.4

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

View this Interactive Content online

Exercise 6.3.5

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

c. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Answer available at the end of this section.

Interactive Exercise 6.3.5

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

View this Interactive Content online

Exercise 6.3.6

The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. We are interested in the length of time a CD player lasts.

a. A CD player is guaranteed for three years. Find the probability that a CD player will break down during the guarantee period.

b. Find the probability that a CD player will last between 2.8 and six years.

c. Find the 70th percentile of the distribution for the time a CD player lasts.

Answer

a. 19.87%
b. 76.94%
c. 4.78 years.

Interactive Exercise 6.3.6

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

View this Interactive Content online

Answers

Exercise 6.3.3

92.63

Exercise 6.3.4

a. Let $X =$ the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < X < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < X < 2.75) = 0.5886$

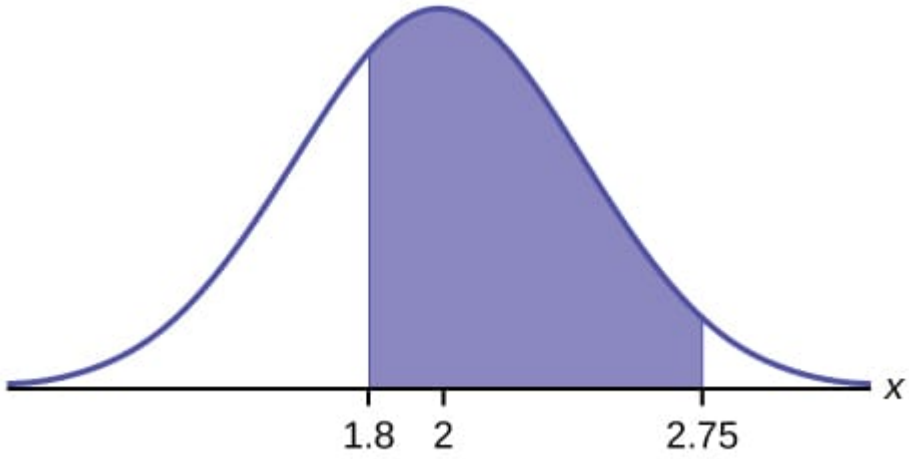


Figure 6.3.5.

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(X < k) = 0.25$.

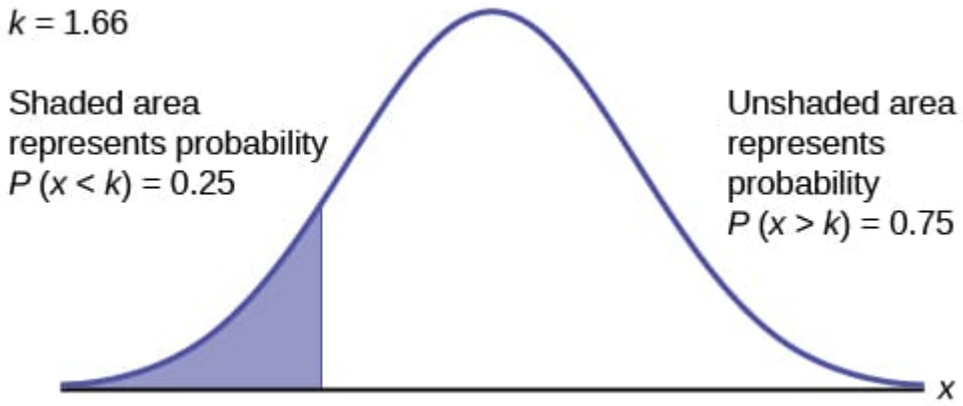


Figure 6.3.6.

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Exercise 6.3.5

a. 0.8186

b. 0.8413

c. The 80th percentile is 48.6 years meaning 80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

This page titled 6.3: Normal Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

6.4: Estimating Population Mean

SECTION OBJECTIVES

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

- Calculate and interpret confidence intervals for estimating a population mean.
- Calculate the sample size required to estimate a population mean given a desired confidence level and margin of error.

Introduction to Estimating

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.



Figure 6.4.1. Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

In this section, you will learn to construct and interpret confidence intervals. Throughout the section, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of $\bar{x}$ and s is called a **statistic** and each of $\bar{\mu}$ and σ is called a **parameter**.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1. \tag{6.4.1}$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \tag{6.4.2}$$

and

$$\bar{x} + 0.2 = 2 + 0.2 = 2.2 \tag{6.4.3}$$

We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is (1.8, 2.2). This 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, $\bar{x}$. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean. When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

> Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Confidence Intervals for Population Means

To construct a confidence interval for a single unknown population mean μ, where the population standard deviation is known, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the **error bound for a population mean** (abbreviated *EBM*). The sample mean $\bar{x}$ is the point estimate of the unknown population mean μ.

The confidence interval estimate will have the form:

$$(\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})$$

or, in symbols,

$$(\bar{x} - EBM, \bar{x} + EBM)$$

The **margin of error** (EBM) depends on the confidence level (abbreviated CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL. α is the probability that the interval does not contain the unknown population parameter. Mathematically,

$$\alpha + CL = 1.$$

Example 6.4.1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5: $\bar{x} = 7$ and $EBM = 2.5$

The confidence interval is (7 – 2.5, 7 + 2.5) and calculating the values gives (4.5, 9.5). If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Exercise 6.4.1

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2. What is the confidence interval estimate for the population mean?

Answer available at the end of this section.

Interactive Exercise 6.4.1

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

View this Interactive Content online

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$. To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10$ in both tails, or 5% in each tail, of the normal distribution.

This is a normal distribution curve. The peak of the curve coincides with the point 10 on the horizontal axis. The points 5 and 15 are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 0.90.

Figure 6.4.1

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the *z*-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is

$$\frac{\sigma}{\sqrt{n}}$$

This fraction is commonly called the "standard error of the mean" to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N(\mu_X, \frac{\sigma}{\sqrt{n}})$.
- When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the *z*-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95, \alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

using a calculator, computer or a standard normal probability table.

Which Calculator to Use?

There are many online calculators that can be used to compute z_α's. For example, you can use this one:

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

View this Interactive Content online

Also, you are encouraged to ask your instructor about which calculator is allowed/recommended for this course.

Example 6.4.2

Use the calculator provided above to verify the following probability statements:

$$z_{0.025} = 1.960$$

$$z_{0.05} = 1.645$$

$$z_{0.005} = 2.576$$

Exercise 6.4.2

Find $z_{0.075}$.

Answer

1.44

Interactive Exercise 6.4.2

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

View this Interactive Content online

Calculating the Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$EBM = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Which Calculator to Use?

There are many online calculators that can be used to compute z_α's. For example, you can use this one:

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

View this Interactive Content online

Also, you are encouraged to ask your instructor about which calculator is allowed/recommended for this course.

Example 6.4.3

Use the calculator provided above to verify the following statements:

When $\alpha = 0.1, n = 50, \sigma = 2.5$ the EBM is 0.5798

When $\alpha = 0.01, n = 75, \sigma = 13.2$ the EBM is 3.9324

When $\alpha = 0.05, n = 10, \sigma = 0.7$ the EBM is 0.4339

Exercise 6.4.3

Find EBM when $\alpha = 0.07, n = 32, \sigma = 6.7$.

Answer

2.1438

Interactive Exercise 6.4.3

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

View this Interactive Content online

Constructing the Confidence Interval

The confidence interval estimate has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

This is a normal distribution curve. The peak of the curve coincides with the point x-bar on the horizontal axis. The points x-bar - EBM and x-bar + EBM are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 1 - a and represents the confidence level. Each unshaded tail has area a/2.

Figure 8.2.2.

Which Calculator to Use?

There are many online calculators that can be used to compute the confidence intervals. For example, you can use this one:

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

View this Interactive Content online

Also, you are encouraged to ask your instructor about which calculator is allowed/recommended for this course.

Example 6.4.4

Use the calculator provided above to verify the following:

Confidence Level (%): 90

Sample Size: 47

Sample Mean: 181.7

Population Standard Deviation: 18.9

90% Confidence Interval: $(177.1654, 186.2346)$

Exercise 6.4.4

Find a 95% confidence interval when the sample size is 30, the sample mean is 60, and the population standard deviation is 5.

Answer

(58.2108, 61.7892)

Interactive Exercise 6.4.4

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

View this Interactive Content online

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example 6.4.5

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z\frac{\alpha}{2}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\sigma = 3; n = 36$$

The confidence level is 90% (CL = 0.90)

$$CL = 0.90$$

so

$$\alpha = 1–CL = 1–0.90 = 0.10$$

$$\frac{\alpha}{2} = 0.05 z\frac{\alpha}{2} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1 - 0.05 = 0.95$.

$$z\frac{\alpha}{2} = z_{0.05} = 1.645$$

using $\text{invNorm}(0.95, 0, 1)$ on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$$

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is **(67.1775, 68.8225).**

Interpretation: We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level: Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Exercise 6.4.5

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes. Find a 90% confidence interval estimate for the population mean delivery time.

Answer available at the end of this section.

Interactive Exercise 6.4.5

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

View this Interactive Content online

Example 6.4.6: Specific Absorption Rate

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution

To find the confidence interval, start by finding the point estimate: the sample mean.

$$\bar{x} = 1.024$$

Next, find the EBM. Because you are creating a 98% confidence interval, $CL = 0.98$.

This is a normal distribution curve. The point z0.01 is labeled at the right edge of the curve and the region to the right of this point is shaded. The area of this shaded region equals 0.01. The unshaded area equals 0.99.

Figure 8.2.3.

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$\bar{x} - EBM = 1.024–0.1431 = 0.8809$

$\bar{x} - EBM = 1.024–0.1431 = 1.1671$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Exercise 6.4.6

Table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Answer

$$\bar{x} = 0.940$$

$$\frac{\alpha}{2} = \frac{1 - CL}{2} = \frac{1 - 0.93}{2} = 0.035$$

$$z_{0.035} = 1.812$$

$$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$$

$$\bar{x} - EBM = 0.940 - 0.1365 = 0.8035$$

$$\bar{x} + EBM = 0.940 + 0.1365 = 1.0765$$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Interactive Content

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

View this Interactive Content online

Notice the difference in the confidence intervals calculated in Example and the following Exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example 6.4.4

Suppose we change the original problem in Example by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3; n = 36$; The confidence level is 95% (*CL* = 0.95).

$CL = 0.95$ so $\alpha = 1–CL = 1–0.95 = 0.05$

$$\frac{\alpha}{2} = 0.025 z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1–0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\bar{x}–EBM = 68–0.98 = 67.02$$

$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the EBM is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

Part (a) shows a normal distribution curve. A central region with area equal to 0.90 is shaded. Each unshaded tail of the curve has area equal to 0.05. Part (b) shows a normal distribution curve. A central region with area equal to 0.95 is shaded. Each unshaded tail of the curve has area equal to 0.025.

Figure 8.2.4.

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population mean when the population standard deviation is known is

$$EBM = \left(z_{\frac{a}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n. In the formulas, z is $z_{\frac{a}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 6.4.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

Solution

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2}$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Exercise 6.4.7

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Answer available at the end of this section.

Interactive Exercise 6.4.7

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

View this Interactive Content online

Answers

Exercise 6.4.1

(11.8, 18.2)

Exercise 6.4.5

(34.1347, 37.8653)

Exercise 6.4.7

35 students

This page titled 6.4: Estimating Population Mean is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 12: Supplementary notes on matrices and linear systems by Isaiah Lankham, Bruno Nachtergaele, and Anne Schilling has no license indicated.

6.5: Estimating Population Proportion

SECTION OBJECTIVES

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

- Calculate and interpret confidence intervals for estimating a population proportion.
- Calculate the sample size required to estimate a population proportion given a desired confidence level and margin of error.

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03,0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the error bound, and the confidence level for a proportion is similar to that for the population mean, but the formulas are different. How do you know you are dealing with a proportion problem? There is no mention of a mean or average!

To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable $\hat{P}$ (read "P hat") is that proportion,

$$\hat{P} = \frac{X}{n}$$

When n is large and p is not close to zero or one, we can use the normal distribution to approximate the number of successes.

$$X \sim N(np, \sqrt{npq})$$

If we divide the random variable, the mean, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion as the number of successes divided by n.)

$$\frac{X}{n} = \hat{P} \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Using algebra to simplify:

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

$\hat{P}$ follows a normal distribution for proportions:

$$\frac{X}{n} = \hat{P} \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Calculating the Error Bound

The error bound (EBP) for a proportion is

$$EBP = (z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

where $\hat{q} = 1 - \hat{p}$.

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is

$$\sqrt{\frac{pq}{n}}.$$

However, in the error bound formula, we use

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

as the standard deviation, instead of

$$\sqrt{\frac{pq}{n}}.$$

In the error bound formula, the sample proportions $\hat{p}$ and $\hat{q}$ are estimates of the unknown population proportions *p* and *q*. The estimated proportions $\hat{p}$ and $\hat{q}$ are used because p and q are not known. The sample proportions $\hat{p}$ and $\hat{q}$ are calculated from the data: $\hat{p}$ is the estimated proportion of successes, and $\hat{q}$ is the estimated proportion of failures.

Which Calculator to Use?

There are many online calculators that can be used to compute the Margin of Error. For example, you can use this one:

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

View this Interactive Content online

Also, you are encouraged to ask your instructor about which calculator is allowed/recommended for this course.

Example 6.5.1

Use the calculator provided above to verify the following statements:

When $\alpha = 0.1, n = 200, \hat{p} = 0.43$ the EBP is 0.0577

When $\alpha = 0.05, n = 100, \hat{p} = 0.81$ the EBP is 0.0768

When $\alpha = 0.01, n = 250, \hat{p} = 0.57$ the EBP is 0.0806

Exercise 6.5.1

Find EBP when $\alpha = 0.07, n = 168, \hat{p} = 0.82$.

Answer

0.0806

Interactive Exercise 6.5.1

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

View this Interactive Content online

Constructing the Confidence Interval

The confidence interval has the form

$$(\hat{p}–EBP, \hat{p} + EBP).$$

where

- EBP is error bound for the proportion.
- $\hat{p} = \frac{x}{n}$
- $\hat{p} =$ the estimated proportion of successes (\hat{p} is a point estimate for *p*, the true proportion.)
- $x =$ the number of successes
- $n =$ the size of the sample

The confidence interval can be used only if the number of successes $n\hat{p}$ and the number of failures $n\hat{q}$ are both greater than five.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

This is a normal distribution curve. The peak of the curve coincides with the point x-bar on the horizontal axis. The points x-bar - EBM and x-bar + EBM are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 1 - a and represents the confidence level. Each unshaded tail has area a/2.

Figure 8.2.2.

Which Calculator to Use?

There are many online calculators that can be used to compute the confidence intervals. For example, you can use this one:

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

View this Interactive Content online

Also, you are encouraged to ask your instructor about which calculator is allowed/recommended for this course.

Example 6.5.2

Use the calculator provided above to verify the following:

Confidence Level (%): 95

Sample Size: 197

Number of Successes: 61

95% Confidence Interval: $(0.2450, 0.3742)$

Exercise 6.5.2

Find a 90% confidence interval when the sample size is 250 and the number of successes is 85.

Answer

(0.2907, 0.3893)

Interactive Exercise 6.5.2

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

View this Interactive Content online

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population proportion**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population proportoin (include the context of the problem) is between ___ and ___ ."

Example 6.5.3

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution

To calculate the confidence interval, you must find $\hat{p}$, $\hat{q}$, and EBP.

- $n = 500$
- $x =$ the number of successes $= 421$

$$\hat{p} = \frac{x}{n} = \frac{421}{500} = 0.842$$

- $\hat{p} = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$\hat{q} = 1–\hat{p} = 1–0.842 = 0.158$$

Since $CL = 0.95$, then

$$\alpha = 1–CL = 1–0.95 = 0.05 \left(\frac{\alpha}{2}\right) = 0.025.$$

Then

$$z\,\frac{\alpha}{2} = z_{0.025=1.96}$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(z\,\frac{\alpha}{2}\right)\sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$\hat{p}–EBP = 0.842–0.032 = 0.81$$

$$\hat{p} + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(\hat{p}–EBP, \hat{p} + EBP) = (0.810, 0.874)$.

Interpretation: We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level: Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Exercise 6.5.3

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Answer available at the end of this section.

Interactive Exercise 6.5.3.1

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

View this Interactive Content online

Interactive Exercise 6.5.3.2

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

View this Interactive Content online

Example 6.5.4

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Answer

$x = 300$ and $n = 500$

$$\hat{p} = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then

$$\alpha = 1\text{–}CL = 1\text{–}0.90 = 0.10 \left(\frac{\alpha}{2}\right) = 0.05 \tag{6.5.1}$$

$$z\,\frac{\alpha}{2} = z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z\,\frac{\alpha}{2}\right)\sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.645)\sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$\hat{p}\text{–}EBP = 0.60 - 0.036 = 0.564$$

$$\hat{p} + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(\hat{p}\text{–}EBP, \hat{p} + EBP) = (0.564, 0.636)$.

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level: Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Exercise 6.5.4

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.
b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Answer available at the end of this section.

Interactive Exercise 6.5.4.1

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

View this Interactive Content online

Interactive Exercise 6.5.4.2

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

View this Interactive Content online

Example 6.5.5

A random sample of 29 statistics students was asked: "Have you smoked a cigarette in the past week?" Eight students reported smoking within the past week. Find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution

Eight students out of 29 reported smoking within the past week, so $x = 8$ and $n = 29$.

$$\hat{p} = \frac{x}{n} = \frac{8}{29} \approx 0.276$$

$$\hat{q} = 1–\hat{p} = 1–0.276 = 0.724$$

Since $CL = 0.95$, we know $\alpha = 1–0.95 = 0.05$ and $\frac{\alpha}{2} = 0.025$.

$$z_{0.025} = 1.96$$

$$EPB = \left(z\,\frac{\alpha}{2} \right) \sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.96)\sqrt{\frac{0.276(0.724)}{29}} \approx 0.163$$

$$\hat{p}–EPB = 0.276–0.163 = 0.113$$

$$\hat{p} + EPB = 0.276 + 0.163 = 0.439$$

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Exercise 6.5.5

Out of a random sample of 69 freshmen at State University, 33 students have declared a major. Find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution

We have $x = 33$ and $n = 69$.

$$\hat{p} = 33/69 \approx 0.478$$

$$\hat{q} = 1–\hat{p} = 1–0.478 = 0.522$$

Since $CL = 0.96$, we know $\alpha = 1–0.96 = 0.04$ and $\frac{\alpha}{2} = 0.02$.

$$z_{0.02} = 2.054$$

$$EPB = \left(z\,\frac{\alpha}{2} \right) \sqrt{\frac{p'q'}{n}} = (2.054) \left(\sqrt{\frac{(0.478)(0.522)}{69}} \right) - 0.124$$

$$\hat{p}–EPB = 0.478–0.124 = 0.354$$

$$\hat{p} + EPB = 0.478 + 0.124 = 0.602$$

We are 96% confident that between 35.4% and 60.2% of all freshmen at State U have declared a major.

Interactive Exercise 6.5.5.1

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

View this Interactive Content online

Interactive Exercise 6.5.5.2

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

View this Interactive Content online

Example 6.5.6

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the “plus four” method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution A

Using "plus-four," we have $x = 13 + 2 = 15$ and $n = 50 + 4 = 54$.

$$p' = \frac{15}{54} \approx 0.278$$

$$q' = 1 - p' = 1 - 0.241 = 0.722$$

Since $CL = 0.90$, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$.

$$z_{0.05} = 1.645$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)(0.722)}{54}}\right) \approx 0.100$$

$$p' - EPB = 0.278 - 0.100 = 0.178$$

$$p' + EPB = 0.278 + 0.100 = 0.378$$

We are 90% confident that between 17.8% and 37.8% of all teens would report having more than 500 friends on Facebook.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 15.
Arrow down to n and enter 54.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.178, 0.378).

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population proportion is

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

Solving for n gives you an equation for the sample size.

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (\hat{p}\hat{q})}{EBP^2}$$

Example 6.5.6

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Answer

From the problem, we know that $\mathbf{EBP = 0.03}$ (3%=0.03) and $z_{\frac{\alpha}{2}}$ $z_{0.05} = 1.645$ because the confidence level is 90%.

However, in order to find n, we need to know the estimated (sample) proportion $\hat{p}$. Remember that $\hat{q} = 1–\hat{p}$. But, we do not know $\hat{p}$ yet. Since we multiply $\hat{p}$ and $\hat{q}$ together, we make them both equal to 0.5 because $\hat{p}\hat{q} = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2\hat{p}\hat{q}}{EBP^2}$$

gives

$$n = \frac{1.645^2(0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Exercise 6.5.6

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Answer available at the end of this section.

Interactive Exercise 6.5.6

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

View this Interactive Content online

Answers

Exercise 6.5.3

(0.3315, 0.4525)

Exercise 6.5.4

(0.7731, 0.8269); We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.31% and 82.69%

Exercise 6.5.6

271 customers should be surveyed. Check the Real Estate section in your local

This page titled 6.5: Estimating Population Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 11: The Spectral Theorem for normal linear maps by Isaiah Lankham, Bruno Nachtergaele, and Anne Schilling has no license indicated.
- 12: Supplementary notes on matrices and linear systems by Isaiah Lankham, Bruno Nachtergaele, and Anne Schilling has no license indicated.

7: MODELING

Close message

Learn more

Welcome to your new guide!

Click on "New" to build out your guide.

- Topic hierarchy

7.1: Introduction to Modeling

7.2: Modeling with Linear Functions

7.3: Fitting Linear Models to Data

7.4: Modeling with Exponential Functions

7.5: Fitting Exponential Models to Data

7.6: Putting It All Together

7.7: Modeling with Quadratic Functions

7.8: Scatter Plots

This page titled 7: Modeling is shared under a license and was authored, remixed, and/or curated by LibreTexts.

7.1: Introduction to Modeling

Learning Objectives

Students will be able to:

- Express the relationship between the input and output variable for a function
- Use function notation to evaluate a function for a given input
- Represent a function as a table, graph, or formula

in order to apply mathematical modeling to solve real-world applications.

In this section, we will introduce the basics for creating mathematical models which enable us to make predictions and understand the relationships that exist between two different factors called variables. We will first discuss functions which give us the tools needed to describe relationships between two variables, and we will then introduce key concepts such as correlation which enable us to create our models.

Functions

Given two sets A and B a set with elements that are ordered pairs (x, y) where x is an element of A and y is an element of B, is a relation from A to B. A relation from A to B defines a relationship between those two sets. A function is a special type of relation in which each element of the first set is related to exactly one element of the second set. The element of the first set is called the ***input***; the element of the second set is called the ***output***. Functions are used all the time in mathematics to describe relationships between two sets. For any function, when we know the input, the output is determined, so we say that the output is a function of the input. For example, the area of a square is determined by its side length, so we say that the area (the output) is a function of its side length (the input). The velocity of a ball thrown in the air can be described as a function of the amount of time the ball is in the air. The cost of mailing a package is a function of the weight of the package. Since functions have so many uses, it is important to have precise definitions and terminology to study them.



Figure 7.1.1: A function can be visualized as an input/output device

Definition: Functions

A *function* f consists of a set of inputs, a set of outputs, and a rule for assigning each input to exactly one output. The set of inputs is called the *domain* of the function. The set of outputs is called the *range* of the *function*.

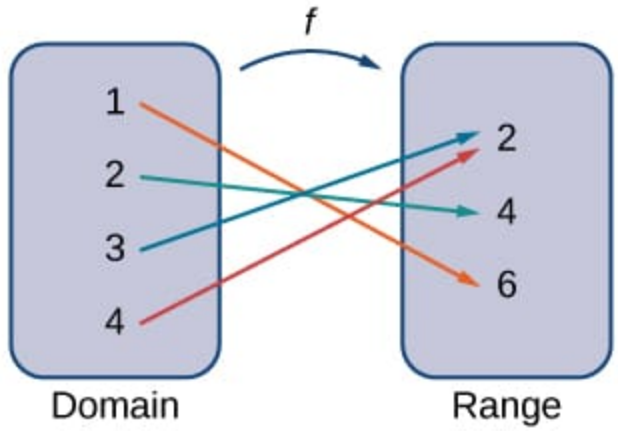


Figure 7.1.2: A function maps every element in the domain to exactly one element in the range. Although each input can be sent to only one output, two different inputs can be sent to the same output.

For example, consider the function f, where the domain is the set of all real numbers and the rule is to square the input. Then, the input $x = 3$ is assigned to the output $3^2 = 9$.

Since every nonnegative real number has a real-value square root, every nonnegative number is an element of the range of this function. Since there is no real number with a square that is negative, the negative real numbers are not elements of the range. We conclude that the range is the set of nonnegative real numbers.

For a general function f with domain D, we often use x to denote the input and y to denote the output associated with x. When doing so, we refer to x as the **independent variable** and y as the **dependent variable**, because it depends on x. Using function notation, we write $y = f(x)$, and we read this equation as "y equals f of x." For the squaring function described earlier, we write $f(x) = x^2$.

The concept of a function can be visualized using Figures 7.1.1 - 7.1.3.

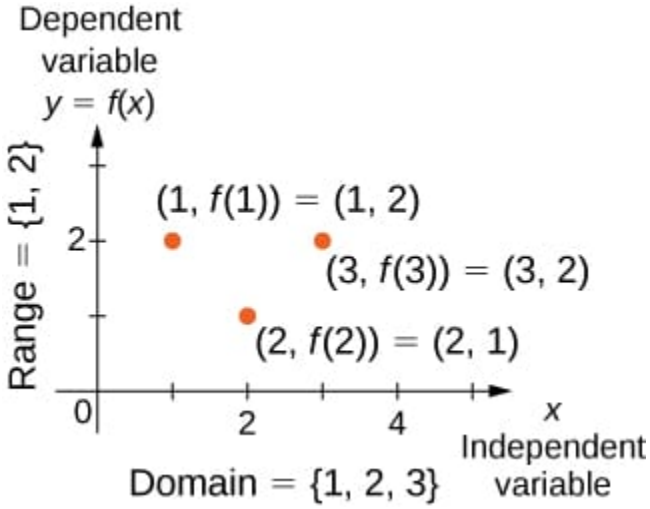


Figure 7.1.3: In this case, a graph of a function f has a domain of $\{1, 2, 3\}$ and a range of $\{1, 2\}$. The independent variable is x and the dependent variable is y.

We can also visualize a function by plotting points (x, y) in the coordinate plane where $y = f(x)$. The graph of a function is the set of all these points. For example, consider the function f, where the domain is the set $D = \{1, 2, 3\}$ and the rule is $f(x) = 3 - x$. In Figure 7.1.4, we plot a graph of this function.

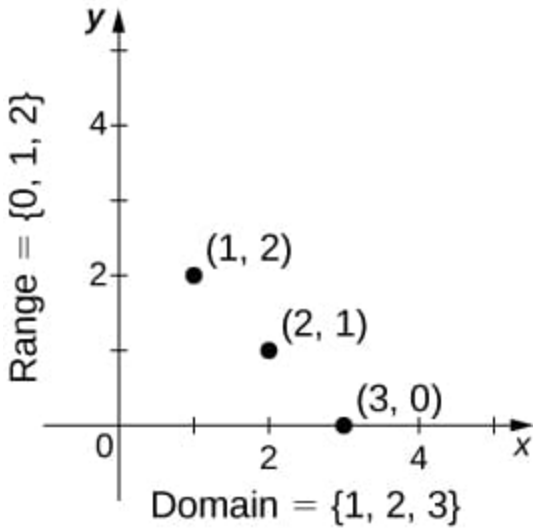


Figure 7.1.4: Here we see a graph of the function f with domain $\{1, 2, 3\}$ and rule $f(x) = 3 - x$. The graph consists of the points $(x, f(x))$ for all x in the domain.

Every function has a domain. However, sometimes a function is described by an equation, as in $f(x) = x^2$, with no specific domain given. In this case, the domain is taken to be the set of all real numbers x for which $f(x)$ is a real number. For example, since any real number can be squared, if no other domain is specified, we consider the domain of $f(x) = x^2$ to be the set of all real numbers. On the other hand, the square root function $f(x) = \sqrt{x}$ only gives a real output if x is nonnegative. Therefore, the domain of the function $f(x) = \sqrt{x}$ is the set of nonnegative real numbers, sometimes called the natural domain.

For the functions $f(x) = x^2$ and $f(x) = \sqrt{x}$, the domains are sets with an infinite number of elements. Clearly we cannot list all these elements. When describing a set with an infinite number of elements, it is often helpful to use set-builder or interval notation. When using set-builder notation to describe a subset of all real numbers, denoted R, we write

$$\{x \,|\, x \ has \ some \ property\}. \tag{7.1.1}$$

We read this as the set of real numbers x such that x has some property. For example, if we were interested in the set of real numbers that are greater than one but less than five, we could denote this set using set-builder notation by writing

$$\{x \,|\, 1 < x < 5\}. \tag{7.1.2}$$

A set such as this, which contains all numbers greater than a and less than b, can also be denoted using the *interval notation* (a, b). Therefore,

$$(1, 5) = \{x \,|\, 1 < x < 5\}. \tag{7.1.3}$$

The numbers 1 and 5 are called the endpoints of this set. If we want to consider the set that includes the endpoints, we would denote this set by writing

$$[1, 5] = \{x \,|\, 1 < x < 5\}. \tag{7.1.4}$$

We can use similar notation if we want to include one of the endpoints, but not the other. To denote the set of nonnegative real numbers, we would use the set-builder notation

$$\{x \,|\, x \geq 0\}. \tag{7.1.5}$$

The smallest number in this set is zero, but this set does not have a largest number. Using interval notation, we would use the symbol ∞, which refers to positive infinity, and we would write the set as

$$[0, \infty) = \{x \,|\, x \geq 0\}. \tag{7.1.6}$$

It is important to note that ∞ is not a real number. It is used symbolically here to indicate that this set includes all real numbers greater than or equal to zero. Similarly, if we wanted to describe the set of all nonpositive numbers, we could write

$$(-\infty, 0] = \{x \mid x \leq 0\}. \tag{7.1.7}$$

Here, the notation $-\infty$ refers to negative infinity, and it indicates that we are including all numbers less than or equal to zero, no matter how small. The set

$$(-\infty, \infty) = \{x \mid x \ is \ any \ real \ number\} \tag{7.1.8}$$

refers to the set of all real numbers.Some functions are defined using different equations for different parts of their domain. These types of functions are known as piecewise-defined functions. For example, suppose we want to define a function f with a domain that is the set of all real numbers such that $f(x) = 3x + 1$ for $x \geq 2$ and $f(x) = x^2$ for$x < 2$.We denote this function by writing

$$f(x) = \begin{cases} 3x + 1, & \text{if } x \geq 2 \\ x^2, & \text{if } x < 2 \end{cases} \tag{7.1.9}$$

When evaluating this function for an input x,the equation to use depends on whether $x \geq 2$ or $x < 2$. For example, since $5 > 2$, we use the fact that $f(x) = 3x + 1$ for $x \geq 2$ and see that $f(5) = 3(5) + 1 = 16$. On the other hand, for $x = -1$, we use the fact that $f(x) = x^2$ for $x < 2$ and see that $f(-1) = 1$.

Example 7.1.1: Evaluating Functions

For the function $f(x) = 3x^2 + 2x - 1$, evaluate:

a. $f(-2)$
b. $f(3)$

Solution

Substitute the given value for x in the formula for $f(x)$.

a. $f(-2) = 3(-2)^2 + 2(-2) - 1 = 12 - 4 - 1 = 7$
b. $f(3) = 3(3)^2 + 2(3) - 1 = 18 + 6 - 1 = 23$

Exercise 7.1.2

For $f(x) = x^2 - 3x + 5$, evaluate $f(1)$ and $f(-1)$.

Hint available at the end of this section.

Representing Functions

Typically, a function is represented using one or more of the following tools:

- A table
- A graph
- A formula

We can identify a function in each form, but we can also use them together. For instance, we can plot on a graph the values from a table or create a table from a formula.

Tables

Functions described using a **table of values** arise frequently in real-world applications. Consider the following simple example. We can describe temperature on a given day as a function of time of day. Suppose we record the temperature every hour for a 24-hour period starting at midnight. We let our input variable x be the time after midnight, measured in hours, and the output variable y be the temperature x hours after midnight, measured in degrees Fahrenheit. We record our data in Table 7.1.1.

Table 7.1.1: Temperature as a Function of Time of Day

Hour After Midnight	Temperature(°F)	Hour After Midnight	Temperature(°F)
0	58	12	84
1	54	13	85
2	53	14	85
3	52	15	83
4	52	16	82
5	55	17	80
6	60	18	77
7	64	19	74
8	72	20	69
9	75	21	65
10	78	22	60
11	80	23	58

We can see from the table that temperature is a function of time, and the temperature decreases, then increases, and then decreases again. However, we cannot get a clear picture of the behavior of the function without graphing it.

Graphs

Given a function f described by a table, we can provide a visual picture of the function in the form of a graph. Graphing the temperatures listed in Table 7.1.1 can give us a better idea of their fluctuation throughout the day. Figure 7.1.5 shows the plot of the temperature function.

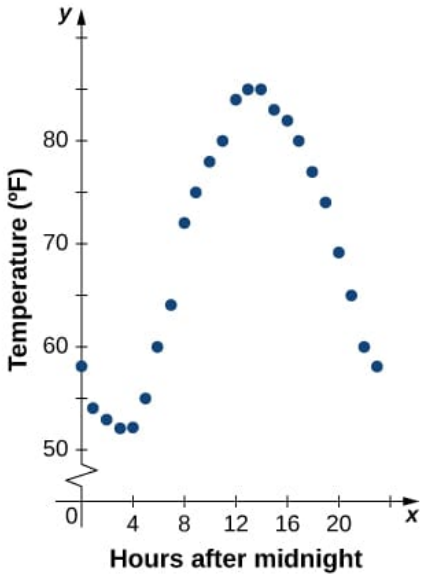


Figure 7.1.5: The graph of the data from Table 7.1.1 shows temperature as a function of time.

From the points plotted on the graph in Figure 7.1.5, we can visualize the general shape of the graph. It is often useful to connect the dots in the graph, which represent the data from the table. In this example, although we cannot make any definitive conclusion regarding what the temperature was at any time for which the temperature was not recorded, given the number of data points collected and the pattern in these points, it is reasonable to suspect that the temperatures at other times followed a similar pattern, as we can see in Figure 7.1.6.

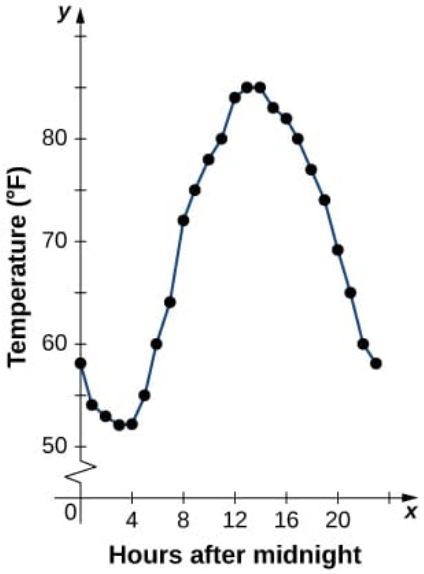


Figure 7.1.6: Connecting the dots in Figure 7.1.5 shows the general pattern of the data.

Algebraic Formulas

Sometimes we are not given the values of a function in table form, rather we are given the values in an explicit formula. Formulas arise in many applications. For example, the area of a circle of radius r is given by the formula $A(r) = \pi r^2$. When an object is thrown upward from the ground with an initial velocity v_0 ft/s, its height above the ground from the time it is thrown until it hits the ground is given by the formula $s(t) = -16t^2 + v_0 t$. When P dollars are invested in an account at an annual interest rate r compounded continuously, the amount of money after t years is given by the formula $A(t) = Pe^{rt}$. Algebraic formulas are important tools to calculate function values. Often we also represent these functions visually in graph form.

Given an algebraic formula for a function f, the graph of f is the set of points $(x, f(x))$, where x is in the domain of f and $f(x)$ is in the range. To graph a function given by a formula, it is helpful to begin by using the formula to create a table of inputs and outputs. If the domain of f consists of an infinite number of values, we cannot list all of them, but because listing some of the inputs and outputs can be very useful, it is often a good way to begin.

When creating a table of inputs and outputs, we typically check to determine whether zero is an output. Those values of x where $f(x) = 0$ are called the zeros of a function. For example, the zeros of $f(x) = x^2 - 4$ are $x = \pm 2$. The zeros determine where the graph of f intersects the x-axis, which gives us more information about the shape of the graph of the function. The graph of a function may never intersect the x-axis, or it may intersect multiple (or even infinitely many) times.

Another point of interest is the y -intercept, if it exists. The y-intercept is given by $(0, f(0))$.

Since a function has exactly one output for each input, the graph of a function can have, at most, one y-intercept. If $x = 0$ is in the domain of a function f, then f has exactly one y-intercept. If $x = 0$ is not in the domain of f, then f has no y-intercept. Similarly, for any real number c, if c is in the domain of f, there is exactly one output $f(c)$, and the line $x = c$ intersects the graph of f exactly once. On the other hand, if c is not in the domain of f, $f(c)$ is not defined and the line $x = c$ does not intersect the graph of f.

Example 7.1.3: Finding the Height of a Free-Falling Object

If a ball is dropped from a height of 100 ft, its height s at time t is given by the function $s(t) = -16t^2 + 100$, where s is measured in feet and t is measured in seconds. The domain is restricted to the interval $[0, c]$, where $t = 0$ is the time when the ball is dropped and $t = c$ is the time when the ball hits the ground.

a. Create a table showing the height s(t) when $t = 0,\ 0.5,\ 1,\ 1.5,\ 2$, and 2.5. Using the data from the table, determine the domain for this function. That is, find the time c when the ball hits the ground.
b. Sketch a graph of s.

Solution

t	0	0.5	1	1.5	2	2.5
$s(t)$	100	96	84	64	36	0

Since the ball hits the ground when $t = 2.5$, the domain of this function is the interval $[0, 2.5]$.

2.

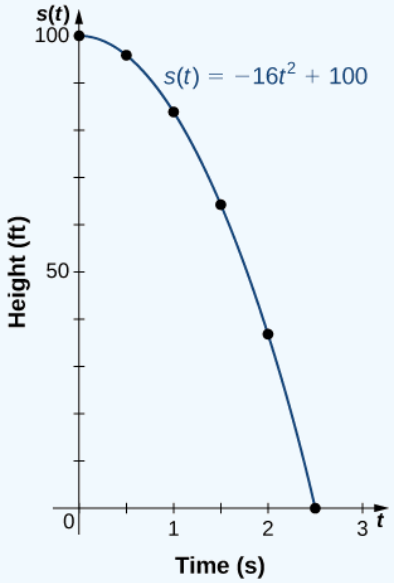


Figure 7.1.8, the values of $f(x)$ are getting smaller as x is getting larger. A function with this property is said to be decreasing. On the other hand, if the values of $f(x)$ were getting larger as the values of x are getting larger, than that function is said to be increasing.

Answers

Exercise 7.1.2

Substitute 1 and -3 for x in the formula for $f(x)$.

$f(1) = (1)^2 - 3(1) + 5 = 1 - 3 + 5 = 3$ and $f(-1) = (-1)^2 - 3(-1) + 5 = 1 + 3 + 5 = 9$

This page titled 7.1: Introduction to Modeling is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 1.1: Review of Functions by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

7.2: Modeling with Linear Functions

Learning Objectives

- Build linear models from verbal descriptions and two data points
- Solve a system of two linear equations using technology
- Construct a scatter plot to determine if there is a positive, negative, or no linear correlation

to apply mathematical modeling to solve real-world applications.

Emily is a college student who plans to spend a summer in Seattle. She has saved $3,500 for her trip and anticipates spending $400 each week on rent, food, and activities. How can we write a linear model to represent her situation? What would be the x-intercept, and what can she learn from it? To answer these and related questions, we can create a model using a linear function. Models such as this one can be extremely useful for analyzing relationships and making predictions based on those relationships. In this section, we will explore examples of **linear function** models.



Figure 7.2.1: (credit: EEK Photography/Flickr)

Identifying Steps to Model and Solve Problems

When **modeling** scenarios with linear functions and solving problems involving quantities with a **constant rate of change**, we typically follow the same problem strategies that we would use for any type of function. Let's briefly review them:

- Identify changing quantities, and then define descriptive variables to represent those quantities. When appropriate, sketch a picture or define a coordinate system.
- Carefully read the problem to identify important information. Look for information that provides values for the variables or values for parts of the functional model, such as slope and initial value.
- Carefully read the problem to determine what we are trying to find, identify, solve, or interpret.
- Identify a solution pathway from the provided information to what we are trying to find. Often this will involve checking and tracking units, building a table, or even finding a formula for the function being used to model the problem.
- When needed, write a formula for the function.
- Solve or evaluate the function using the formula.
- Reflect on whether your answer is reasonable for the given situation and whether it makes sense mathematically.
- Clearly convey your result using appropriate units, and answer in full sentences when necessary.

Building Linear Models

Now let's take a look at the student in Seattle. In her situation, there are two changing quantities: time and money. The amount of money she has remaining while on vacation depends on how long she stays. We can use this information to define our variables, including units.

- Output: M, money remaining, in dollars
- Input: t, time, in weeks

So, the amount of money remaining depends on the number of weeks: $M(t)$

We can also identify the initial value and the rate of change.

- Initial Value: She saved $3,500, so $3,500 is the initial value for M.
- Rate of Change: She anticipates spending $400 each week, so –$400 per week is the rate of change, or slope.

Notice that the unit of dollars per week matches the unit of our output variable divided by our input variable. Also, because the slope is negative, the linear function is decreasing. This should make sense because she is spending money each week.

The **rate of change is constant**, so we can start with the linear model $M(t) = mt + b$. Then we can substitute the intercept and slope provided.

$$M(t) = mt + b$$

$$M(t) = -400t + 3500$$

Figure 7.2.2

To find the x-intercept, we set the output to zero, and solve for the input.

$$\begin{aligned} 0 &= -400t + 3500 \\ t &= \frac{3500}{400} \\ &= 8.75 \end{aligned}$$

The x-intercept is 8.75 weeks. Because this represents the input value when the output will be zero, we could say that Emily will have no money left after 8.75 weeks.

When modeling any real-life scenario with functions, there is typically a limited domain over which that model will be valid—almost no trend continues indefinitely. Here the domain refers to the number of weeks. In this case, it doesn't make sense to talk about input values less than zero. A negative input value could refer to a number of weeks before she saved $3,500, but the scenario discussed poses the question once she saved $3,500 because this is when her trip and subsequent spending starts. It is also likely that this model is not valid after the x-intercept, unless Emily will use a credit card and goes into debt. The domain represents the set of input values, so the reasonable domain for this function is $0 \leq t \leq 8.75$.

In the above example, we were given a written description of the situation. We followed the steps of modeling a problem to analyze the information. However, the information provided may not always be the same. Sometimes we might be provided with an intercept. Other times we might be provided with an output value. We must be careful to analyze the information we are given, and use it appropriately to build a linear model.

Using a Given Intercept to Build a Model

Some real-world problems provide the y-intercept, which is the constant or initial value. Once the y-intercept is known, the x-intercept can be calculated. Suppose, for example, that Hannah plans to pay off a no-interest loan from her parents. Her loan

balance is $1,000. She plans to pay $250 per month until her balance is $0. The y-intercept is the initial amount of her debt, or $1,000. The rate of change, or slope, is -$250 per month. We can then use the slope-intercept form and the given information to develop a linear model.

$$\begin{aligned} f(x) &= mx + b \\ &= -250x + 1000 \end{aligned}$$

Now we can set the function equal to 0, and solve for x to find the x-intercept.

$$\begin{aligned} 0 &= -250 + 1000 \\ 1000 &= 250x \\ 4 &= x \\ x &= 4 \end{aligned}$$

The x-intercept is the number of months it takes her to reach a balance of $0. The x-intercept is 4 months, so it will take Hannah four months to pay off her loan.

Using a Given Input and Output to Build a Model

Many real-world applications are not as direct as the ones we just considered. Instead they require us to identify some aspect of a linear function. We might sometimes instead be asked to evaluate the linear model at a given input or set the equation of the linear model equal to a specified output.



Given a word problem that includes two pairs of input and output values, use the linear function to solve a problem.

1. Identify the input and output values.
2. Convert the data to two coordinate pairs.
3. Find the slope.
4. Write the linear model.
5. Use the model to make a prediction by evaluating the function at a given x-value.
6. Use the model to identify an x-value that results in a given y-value.
7. Answer the question posed.

Example 7.2.1: Using a Linear Model to Investigate a Town's Population

A town's population has been growing linearly. In 2004 the population was 6,200. By 2009 the population had grown to 8,100. Assume this trend continues.

a. Predict the population in 2013.
b. Identify the year in which the population will reach 15,000.

Solution

The two changing quantities are the population size and time. While we could use the actual year value as the input quantity, doing so tends to lead to very cumbersome equations because the y-intercept would correspond to the year 0, more than 2000 years ago!

To make computation a little nicer, we will define our input as the number of years since 2004:

- Input: t, years since 2004
- Output: $P(t)$, the town's population

To predict the population in 2013 ($t = 9$), we would first need an equation for the population. Likewise, to find when the population would reach 15,000, we would need to solve for the input that would provide an output of 15,000. To write an equation, we need the initial value and the rate of change, or slope.

To determine the rate of change, we will use the change in output per change in input.

$$m = \frac{\text{change in output}}{\text{change in input}} \tag{7.2.1}$$

The problem gives us two input-output pairs. Converting them to match our defined variables, the year 2004 would correspond to $t = 0$, giving the point $(0, 6200)$. Notice that through our clever choice of variable definition, we have "given" ourselves the y-intercept of the function. The year 2009 would correspond to $t = 5$, giving the point $(5, 8100)$.

The two coordinate pairs are $(0, 6200)$ and $(5, 8100)$. Recall that we encountered examples in which we were provided two points earlier in the chapter. We can use these values to calculate the slope.

$$\begin{aligned} m &= \frac{8100 - 6200}{5 - 0} \\ &= \frac{1900}{5} \\ &= 380 \text{ people per year} \end{aligned}$$

We already know the y-intercept of the line, so we can immediately write the equation:

$$P(t) = 380t + 6200 \tag{7.2.2}$$

To predict the population in 2013, we evaluate our function at $t = 9$.

$$\begin{aligned} P(9) &= 380(9) + 6,200 \\ &= 9,620 \end{aligned}$$

If the trend continues, our model predicts a population of 9,620 in 2013.

To find when the population will reach 15,000, we can set $P(t) = 15000$ and solve for t.

$$\begin{aligned} 15000 &= 380t + 6200 \\ 8800 &= 380t \\ t &\approx 23.158 \end{aligned}$$

Our model predicts the population will reach 15,000 in a little more than 23 years after 2004, or somewhere around the year 2027.

Exercise 7.2.1A

A company sells doughnuts. They incur a fixed cost of $25,000 for rent, insurance, and other expenses. It costs $0.25 to produce each doughnut.

a. Write a linear model to represent the cost C of the company as a function of x, the number of doughnuts produced.
b. Find and interpret the y-intercept.

Solution

a. $C(x) = 0.25x + 25,000$ b. The y-intercept is $(0, 25,000)$. If the company does not produce a single doughnut, they still incur a cost of $25,000.

Exercise 7.2.1B

A city's population has been growing linearly. In 2008, the population was 28,200. By 2012, the population was 36,800. Assume this trend continues.

a. Predict the population in 2014.
b. Identify the year in which the population will reach 54,000.

Solution

a. 41,100 b. 2020

Building Systems of Linear Models

The previous section involved only one linear model. Oftentimes, real-world situations will include two or more linear functions which may be modeled with a **system of linear equations**. When solving a system of linear equations, we are looking for points the two lines have in common. Typically, there are three types of answers possible, as shown in Figure 7.2.6.

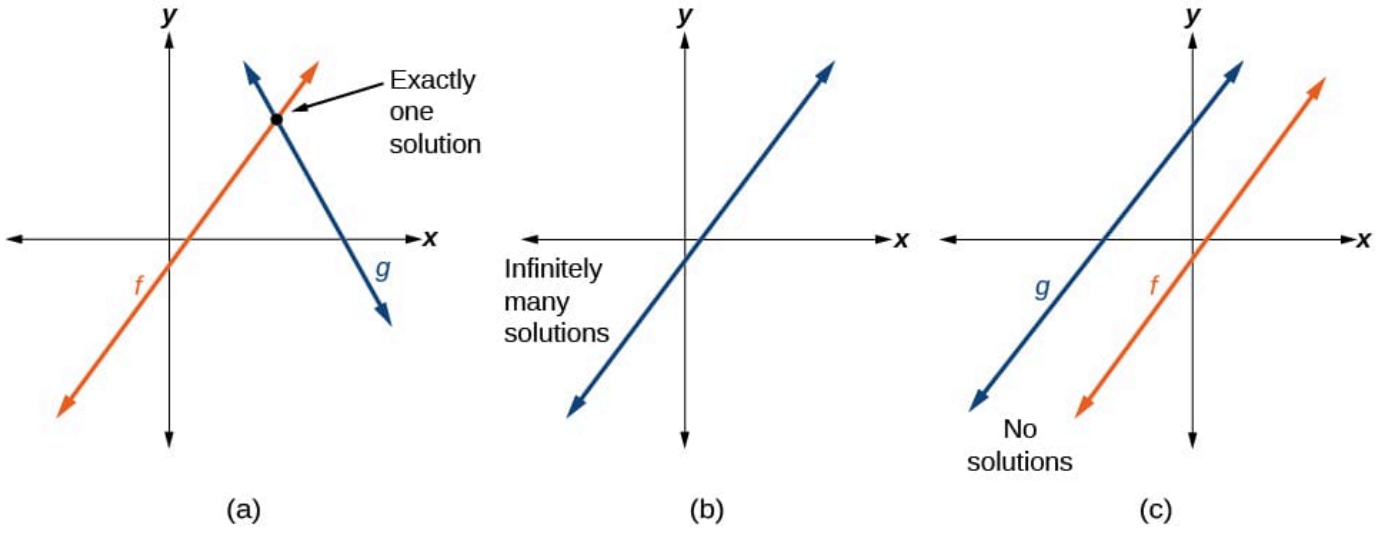


Figure 7.2.3

How To: Given a situation that represents a system of linear equations, write the system of equations and identify the solution.

a. Identify the input and output of each linear model.
b. Identify the slope and y-intercept of each linear model.
c. Find the solution by setting the two linear functions equal to another and solving for x,or find the point of intersection on a graph (this is most easily done with technology such as desmos).

Example 7.2.2: Building a System of Linear Models to Choose a Truck Rental Company

Jamal is choosing between two truck-rental companies. The first, Keep on Trucking, Inc., charges an up-front fee of $20, then 59 cents a mile[1]. The second, Move It Your Way, charges an up-front fee of $16, then 63 cents a mile. When will Keep on Trucking, Inc. be the better choice for Jamal?

Solution

The two important quantities in this problem are the cost and the number of miles driven. Because we have two companies to consider, we will define two functions.

- Input: d, distance driven in miles
- Outputs: $K(d)$: cost, in dollars, for renting from Keep on Trucking

  $M(d)$: cost, in dollars, for renting from Move It Your Way
- Initial Value: Up-front fee: $K(0) = 20$ and $M(0) = 16$

- Rate of Change: $K(d) = \frac{\$0.59}{\text{mile}}$ and $P(d) = \frac{\$0.63}{\text{mile}}$

A linear function is of the form $f(x) = mx + b$. Using the rates of change and initial charges, we can write the equations

$$K(d) = 0.59d + 20$$

$$M(d) = 0.63d + 16$$

Using these equations, we can determine when Keep on Trucking, Inc., will be the better choice. Because all we have to make that decision from is the costs, we are looking for when Move It Your Way, will cost less, or when $K(d) < M(d)$. The solution pathway will lead us to find the equations for the two functions, find the intersection, and then see where the $K(d)$ function is smaller.

These graphs are sketched in Figure 7.2.7, with $K(d)$ in blue.

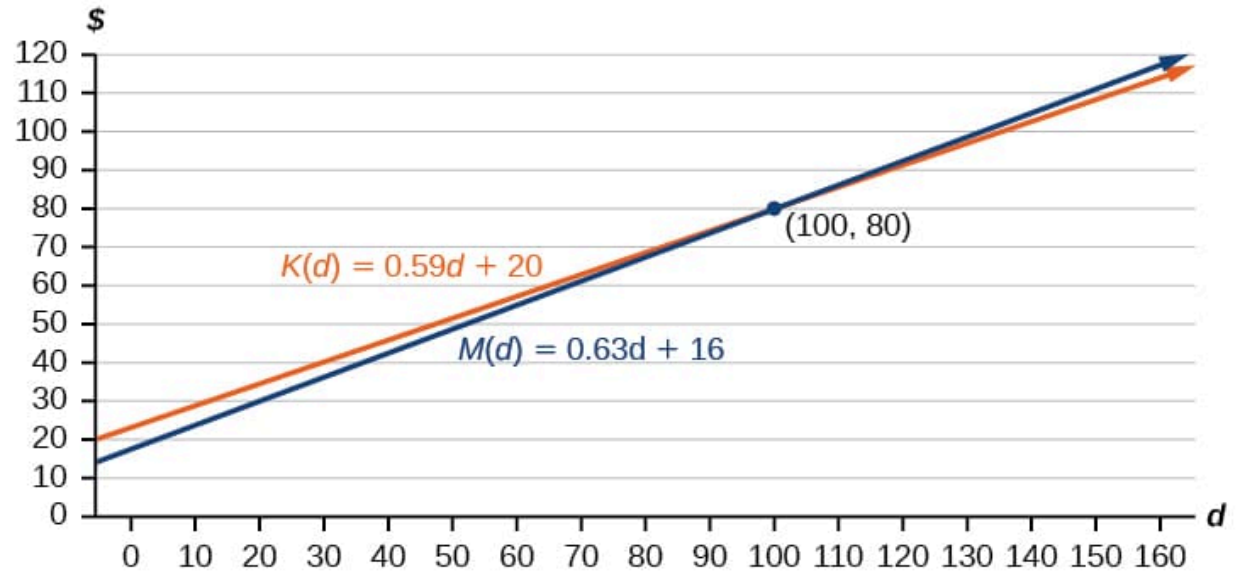


Figure 7.2.4

To find the intersection, we set the equations equal and solve:

$$\begin{aligned} K(d) &= M(d) \\ 0.59d + 20 &= 0.63d + 16 \\ 4 &= 0.04d \\ 100 &= d \\ d &= 100 \end{aligned}$$

This tells us that the cost from the two companies will be the same if 100 miles are driven. Either by looking at the graph, or noting that $K(d)$ is growing at a slower rate, we can conclude that Keep on Trucking, Inc. will be the cheaper price when more than 100 miles are driven, that is $d > 100$.

Please note this problem is easily solved by typing in the equations into desmos and clicking on the point of intersection to highlight the coordinates. You can see the result at https://www.desmos.com/calculator/hlnsxjksmy

Scatter Plots and Correlation

So far, we have been exploring linear relationships where we are given two data points which are used to construct the model. However, we will find that in many real-world applications that we have more than just two data points. In those cases, it's likely that the points don't all perfectly line up which will make constructing a linear model more difficult. It's important to first graph the points on a scatter plot to see if the data roughly follows a linear pattern, which is also known as a linear correlation. A **correlation**

is simply a pattern that connects the input and output variables, and our goal is to investigate if the pattern is linear, and if it is, to eventually come up with an equation that expresses the correlation.

Let's start by exploring scatter plots with an example problem. Please remember that scatter plots are most easily constructed by using technology such as desmos.

Example 7.2.3

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let $x =$ the year and let $y =$ the number of m-commerce users, in millions.

Table 7.2.1: Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

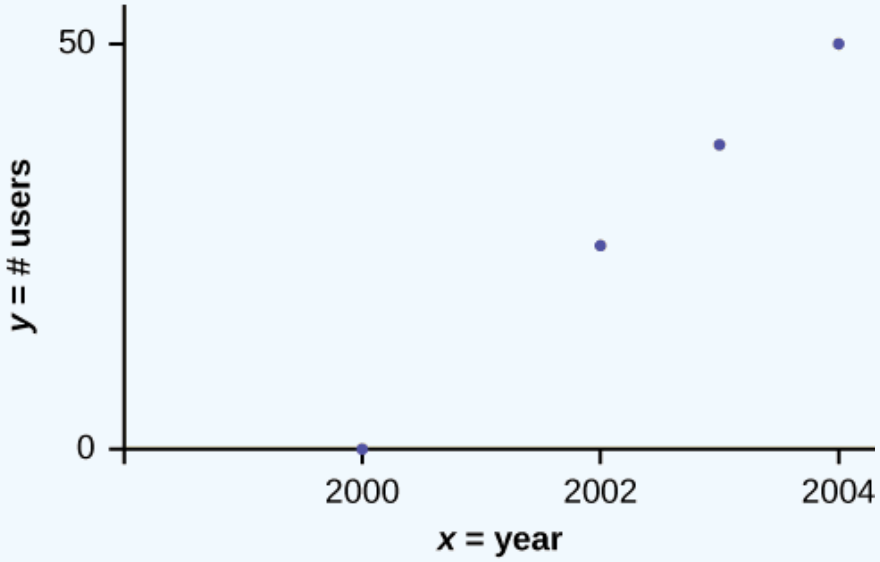


Figure 7.2.5: Scatter plot showing the number of m-commerce users (in millions) by year.

To create a scatter plot

Go to desmos and create a table. From there, fill in the first column with the input values, and the second column with the associated output values. Make sure that the ordered pairs are still together and double-check that your table matches the given information! From there, the graph of the points will automatically generate.

Exercise 7.2.4

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Answer available at the end of this section.

A scatter plot shows the *direction of a relationship* between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the *strength of the relationship* by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatter plot, you want to notice the *overall pattern* and any *deviations* from the pattern. The following scatterplot examples illustrate these concepts.

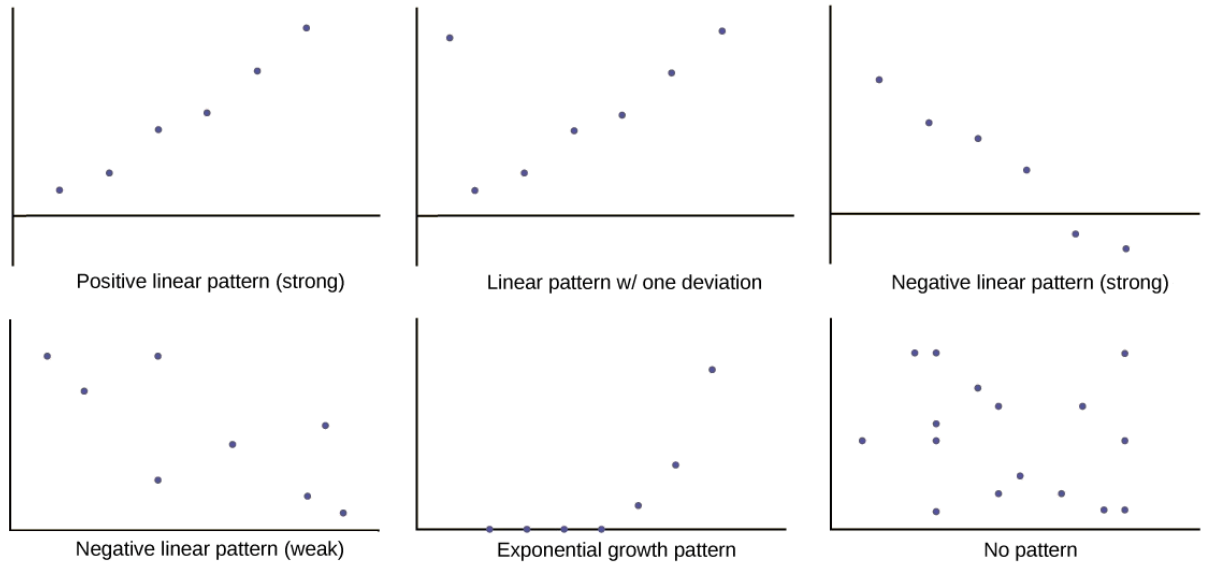


Figure 7.2.7:

In the next section, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear

relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

Answers

Exercise 7.2.4

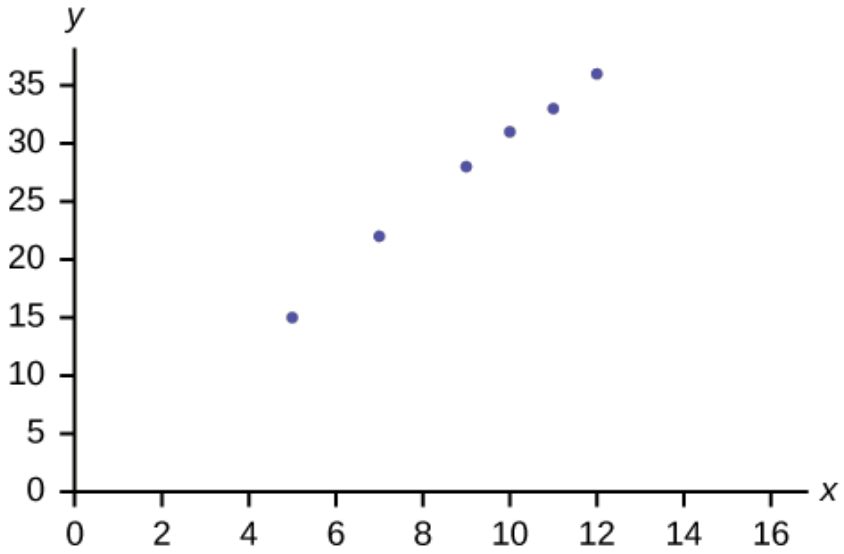


Figure 7.2.6

Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

This page titled 7.2: Modeling with Linear Functions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 2.3: Modeling with Linear Functions by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.
- 1.1: Review of Functions by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

7.3: Fitting Linear Models to Data

Learning Objectives

- Draw and interpret scatter diagrams
- Use a graphing utility to find the line of best fit
- Fit a regression line to a set of data and use the linear model to make predictions

in order to apply mathematical modeling to solve real-world applications.

A professor is attempting to identify trends among final exam scores. His class has a mixture of students, so he wonders if there is any relationship between age and final exam scores. One way for him to analyze the scores is by creating a diagram that relates the age of each student to the exam score received. In this section, we will examine one such diagram known as a scatter plot.

Drawing and Interpreting Scatter Plots

A scatter plot is a graph of plotted points that may show a relationship between two sets of data. If the relationship is from a **linear model**, or a model that is nearly linear, the professor can draw conclusions using his knowledge of linear functions. Figure 7.3.1 shows a sample scatter plot.

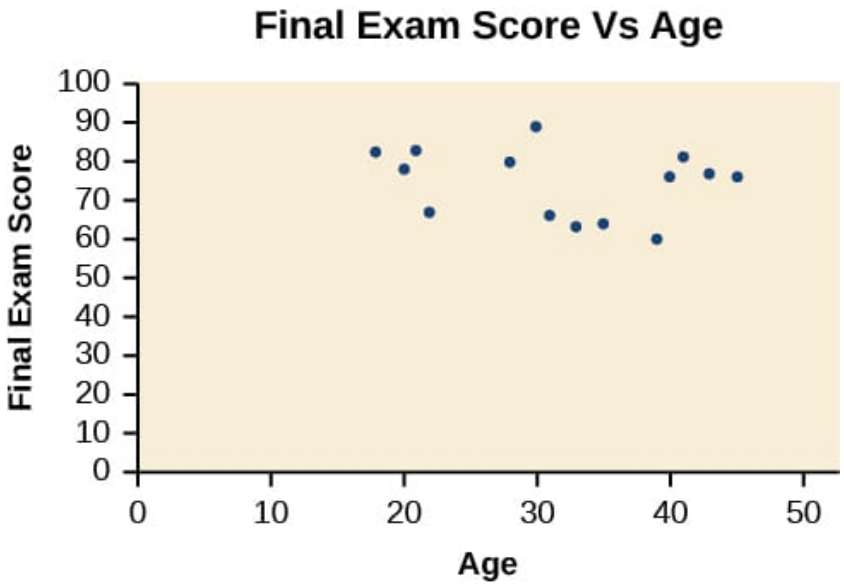


Figure 7.3.1: A scatter plot of age and final exam score variables

Notice this scatter plot does not indicate a **linear relationship**. The points do not appear to follow a trend. In other words, there does not appear to be a relationship between the age of the student and the score on the final exam.

Example 7.3.1: Using a Scatter Plot to Investigate Cricket Chirps

Table shows the number of cricket chirps in 15 seconds, for several different air temperatures, in degrees Fahrenheit[1]. Plot this data, and determine whether the data appears to be linearly related.

Table 7.3.1

Chirps	44	35	20.4	33	31	35	18.5	37	26

Temperature	80.5	70.5	57	66	68	72	52	73.5	53

Solution

Plotting this data, as depicted in Figure 7.3.2 suggests that there may be a trend. We can see from the trend in the data that the number of chirps increases as the temperature increases. The trend appears to be roughly linear, though certainly not perfectly so.

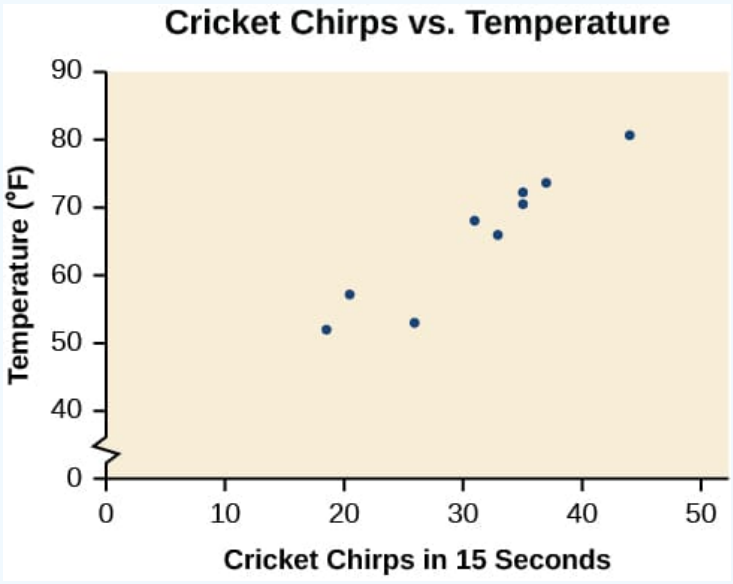


Figure 7.3.2: The line regression is generally positive.

Finding the Line of Best Fit

Once we recognize a need for a linear function to model that data, the natural follow-up question is "what is that linear function?" One way to approximate our linear function is to sketch the line that seems to best fit the data. Then we can extend the line until we can verify the y-intercept. We can approximate the slope of the line by extending it until we can estimate the $\frac{\text{rise}}{\text{run}}$.

Example 7.3.2: Finding a Line of Best Fit

Find a linear function that fits the data in Table 7.3.1 by "eyeballing" a line that seems to fit.

Solution

On a graph, we could try sketching a line.

Using the starting and ending points of our hand drawn line, points $(0, 30)$ and $(50, 90)$, this graph has a slope of

$$m = \frac{60}{50} = 1.2 \tag{7.3.1}$$

and a y-intercept at 30. This gives an equation of

$$T(c) = 1.2c + 30 \tag{7.3.2}$$

where c is the number of chirps in 15 seconds, and $T(c)$ is the temperature in degrees Fahrenheit. The resulting equation is represented in Figure 7.3.3.

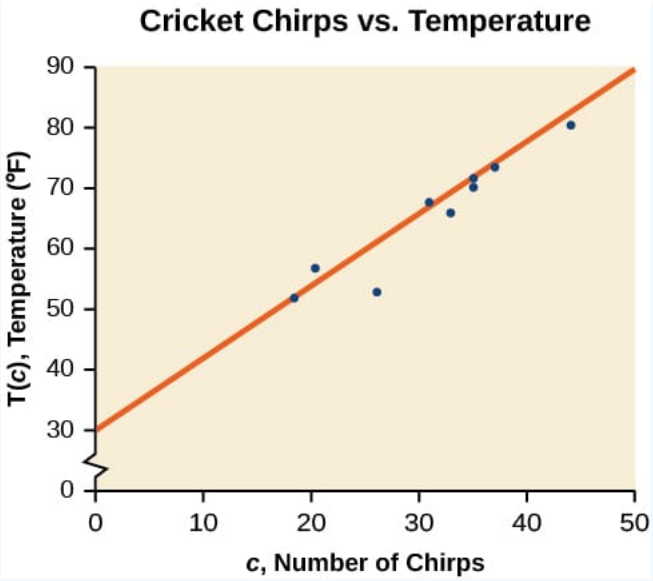


Figure 7.3.3: Scatter plot, showing the line of best fit.

Analysis

This linear equation can then be used to approximate answers to various questions we might ask about the trend.

Recognizing Interpolation or Extrapolation

While the data for most examples does not fall perfectly on the line, the equation is our best guess as to how the relationship will behave outside of the values for which we have data. We use a process known as **interpolation** when we predict a value inside the domain and range of the data. The process of **extrapolation** is used when we predict a value outside the domain and range of the data.

Figure 7.3.4 compares the two processes for the cricket-chirp data addressed in Example 7.3.2. We can see that interpolation would occur if we used our model to predict temperature when the values for chirps are between 18.5 and 44. Extrapolation would occur if we used our model to predict temperature when the values for chirps are less than 18.5 or greater than 44.

There is a difference between making predictions inside the domain and range of values for which we have data and outside that domain and range. Predicting a value outside of the domain and range has its limitations. When our model no longer applies after a certain point, it is sometimes called **model breakdown**. For example, predicting a cost function for a period of two years may involve examining the data where the input is the time in years and the output is the cost. But if we try to extrapolate a cost when $x = 50$, that is in 50 years, the model would not apply because we could not account for factors fifty years in the future.

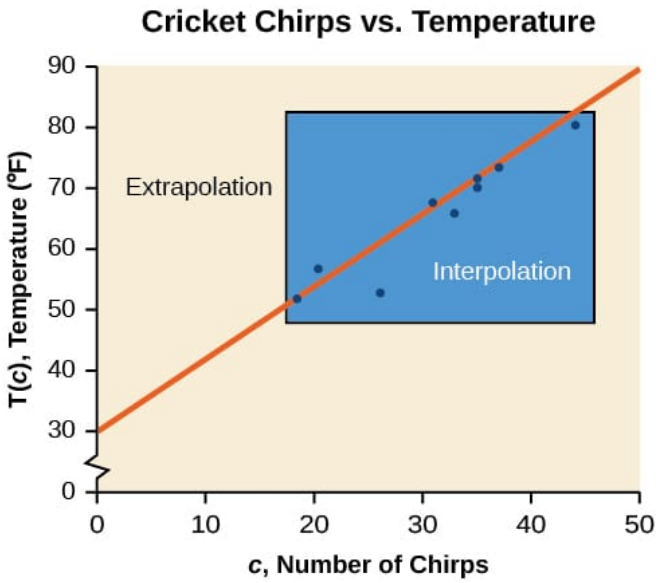


Figure 7.3.4: Interpolation occurs within the domain and range of the provided data whereas extrapolation occurs outside.

Interpolation and Extrapolation

Different methods of making predictions are used to analyze data.

- The method of **extrapolation** involves predicting a value outside the domain and/or range of the data.
- **Model breakdown** occurs at the point when the model no longer applies.

Example 7.3.3: Understanding Interpolation and Extrapolation

Use the cricket data from Table 7.3.1 to answer the following questions:

a. Would predicting the temperature when crickets are chirping 30 times in 15 seconds be interpolation or extrapolation? Make the prediction, and discuss whether it is reasonable.
b. Would predicting the number of chirps crickets will make at 40 degrees be interpolation or extrapolation? Make the prediction, and discuss whether it is reasonable.

Solution

a. The number of chirps in the data provided varied from 18.5 to 44. A prediction at 30 chirps per 15 seconds is inside the domain of our data, so would be interpolation. Using our model:

$$T(30) = 30 + 1.2(30) \tag{7.3.3}$$
$$= 66 \text{ degrees} \tag{7.3.4}$$

Based on the data we have, this value seems reasonable.
b. The temperature values varied from 52 to 80.5. Predicting the number of chirps at 40 degrees is extrapolation because 40 is outside the range of our data. Using our model:

$$40 = 30 + 1.2c \tag{7.3.5}$$
$$10 = 1.2c \tag{7.3.6}$$
$$c \approx 8.33 \tag{7.3.7}$$

We can compare the regions of interpolation and extrapolation using Figure 7.3.5.

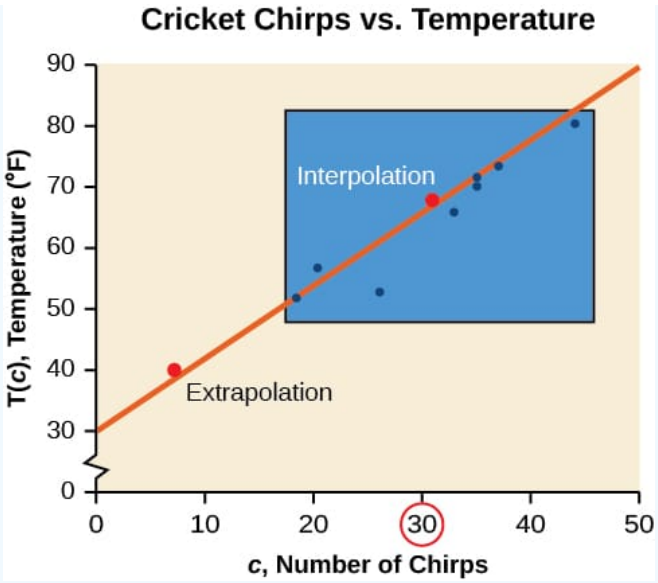


Figure 7.3.5: Scatter plot, showing the line of best fit and where interpolation and extrapolation occurs.

Analysis

Our model predicts the crickets would chirp 8.33 times in 15 seconds. While this might be possible, we have no reason to believe our model is valid outside the domain and range. In fact, generally crickets stop chirping altogether below around 50 degrees.

Exercise 7.3.1

According to the data from Table 7.3.1, what temperature can we predict it is if we counted 20 chirps in 15 seconds?

Solution

54°F

Finding the Line of Best Fit Using a Graphing Utility

While eyeballing a line works reasonably well, there are statistical techniques for fitting a line to data that minimize the differences between the line and data values[2]. One such technique is called **least squares regression** and can be computed by many graphing calculators, spreadsheet software, statistical software, and many web-based calculators such as desmos[3]. Least squares regression is one means to determine the line that best fits the data, and here we will refer to this method as linear regression.

How To: Given data of input and corresponding outputs from a linear function, find the best fit line using linear regression.

1. Enter the input (chirps or c) in List 1 (L1).
2. Enter the output (temperature or T) in List 2 (L2).
3. On a graphing utility, select Linear Regression (LinReg) or if using demos input T ~ mc + b.

Example 7.3.4: Finding a Least Squares Regression Line

Find the least squares regression line using the cricket-chirp data in Table 7.3.1.

Solution

Enter the input (chirps or c) in List 1 (L1).
Enter the output (temperature or T) in List 2 (L2). See Table 7.3.2.

Table 7.3.2

L1	44	35	20.4	33	31	35	18.5	37	26
L2	80.5	70.5	57	66	68	72	52	73.5	53

On a graphing utility, select Linear Regression (LinReg) or if using demos input T ~ mc + b. Using the cricket chirp data from earlier, with technology we obtain the equation:

$$T(c) = 30.281 + 1.143c \tag{7.3.8}$$

Analysis

Notice that this line is quite similar to the equation we "eyeballed" but should fit the data better. Notice also that using this equation would change our prediction for the temperature when hearing 30 chirps in 15 seconds from 66 degrees to:

$$\begin{aligned} T(30) &= 30.281 + 1.143(30) && (7.3.9)\\ &= 64.571 && (7.3.10)\\ &\approx 64.6 \text{ degrees} && (7.3.11) \end{aligned}$$

The graph of the scatter plot with the least squares regression line is shown in Figure 7.3.6 and the result on desmos is shown in Figure 7.3.7. You can access the desmos graph at https://www.desmos.com/calculator/pphrzrssmh.

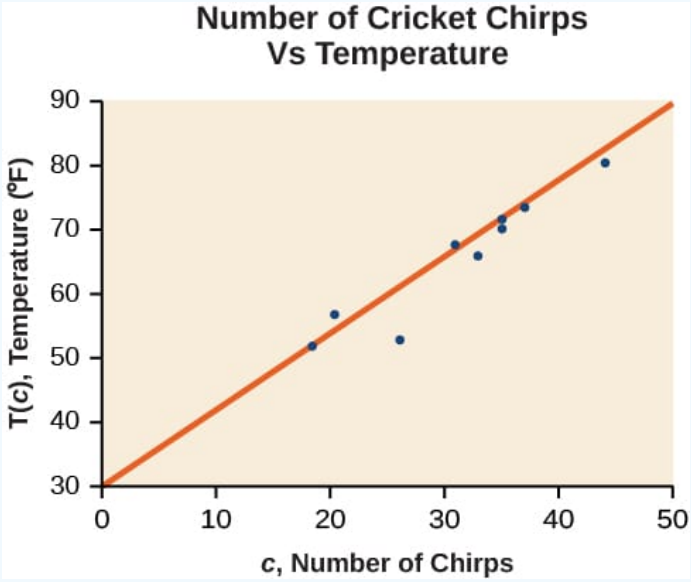


Figure 7.3.6: Scatter plot, showing the line of best fit.

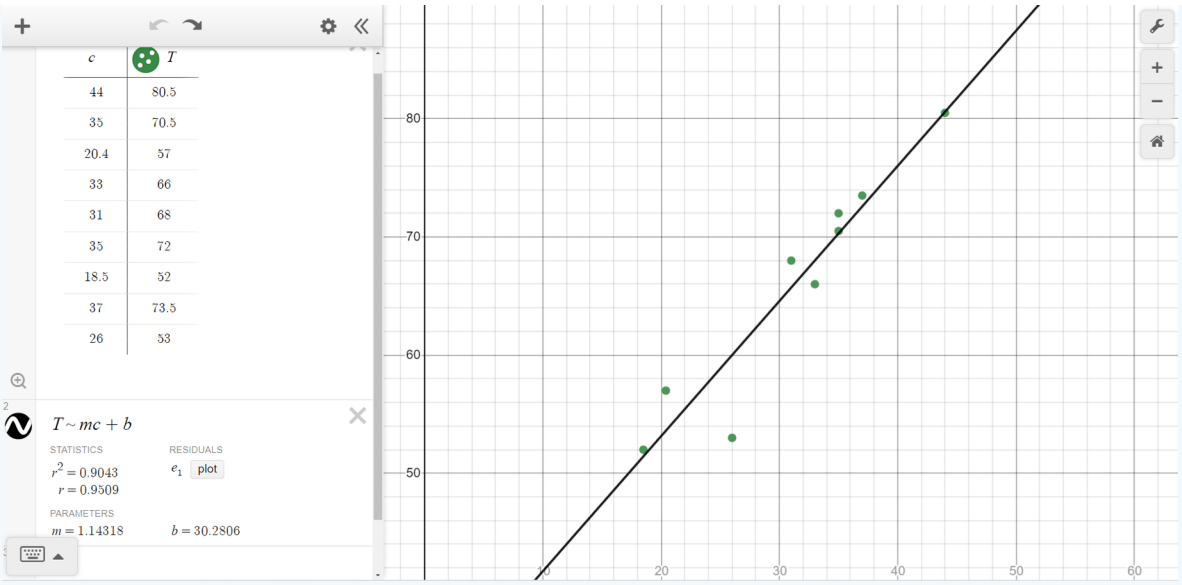


Figure 7.3.7: Scatter plot, showing the line of best fit on desmos.

Will there ever be a case where two different lines will serve as the best fit for the data?

No. There is only one best fit line.

Distinguishing Between Linear and Non-Linear Models

As we saw above with the cricket-chirp model, some data exhibit strong linear trends, but other data, like the final exam scores plotted by age, are clearly nonlinear. Most calculators and computer software can also provide us with the **correlation coefficient**, which is a measure of how closely the line fits the data. Many graphing calculators require the user to turn a "diagnostic on" selection to find the correlation coefficient, which mathematicians label as r. The correlation coefficient provides an easy way to get an idea of how close to a line the data falls.

We should compute the correlation coefficient only for data that follows a linear pattern or to determine the degree to which a data set is linear. If the data exhibits a nonlinear pattern, the correlation coefficient for a linear regression is meaningless. To get a sense for the relationship between the value of r and the graph of the data, Figure 7.3.8 shows some large data sets with their correlation coefficients. Remember, for all plots, the horizontal axis shows the input and the vertical axis shows the output.

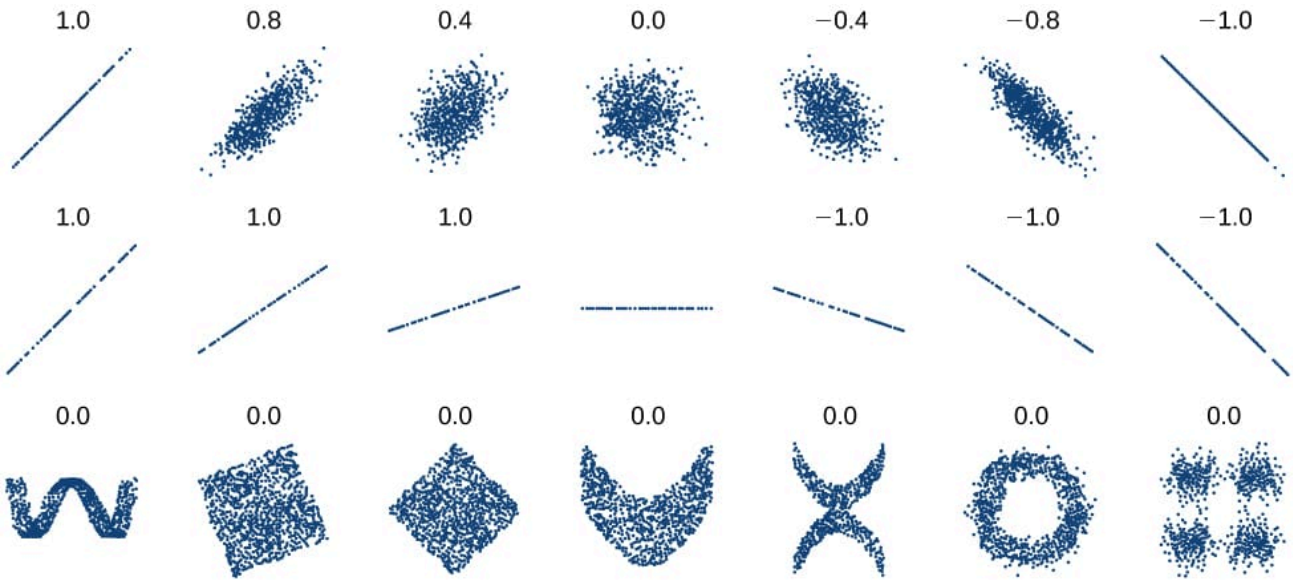


Figure 7.3.8: Plotted data and related correlation coefficients. (credit: "DenisBoigelot," Wikimedia Commons)

Correlation Coefficient

The **correlation coefficient** is a value, r, between –1 and 1.

- $r > 0$ suggests a positive (increasing) relationship
- $r < 0$ suggests a negative (decreasing) relationship
- The closer the value is to 0, the more scattered the data.
- The closer the value is to 1 or –1, the less scattered the data is.

Example 7.3.5: Finding a Correlation Coefficient

Calculate the correlation coefficient for cricket-chirp data in Table 7.3.1.

Solution

Because the data appear to follow a linear pattern, we can use technology to calculate r. Enter the inputs and corresponding outputs and select the Linear Regression. The calculator will also provide you with the correlation coefficient, $r = 0.9509$. This value is very close to 1, which suggests a strong increasing linear relationship.

Note: For some calculators, the Diagnostics must be turned "on" in order to get the correlation coefficient when linear regression is performed: [2nd]>[0]>[alpha][x–1], then scroll to DIAGNOSTICSON.

Predicting with a Regression Line

Once we determine that a set of data is linear using the correlation coefficient, we can use the regression line to make predictions. As we learned above, a regression line is a line that is closest to the data in the scatter plot, which means that only one such line is a best fit for the data.

Example 7.3.6: Using a Regression Line to Make Predictions

Gasoline consumption in the United States has been steadily increasing. Consumption data from 1994 to 2004 is shown in Table 7.3.3. Determine whether the trend is linear, and if so, find a model for the data. Use the model to predict the consumption in 2008.

Table 7.3.3

Year	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04
Consumption (billions of gallons)	113	116	118	119	123	125	126	128	131	133	136

The scatter plot of the data, including the least squares regression line, is shown in Figure 7.3.9.

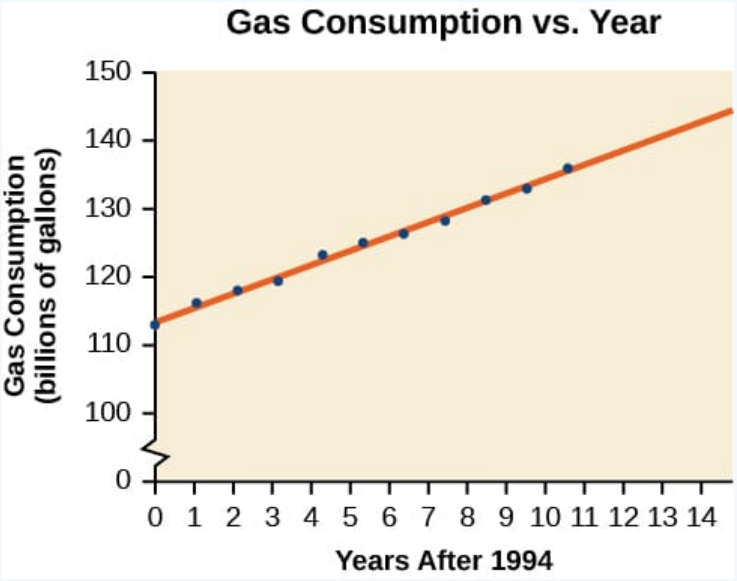


Figure 7.3.9: Scatter plot, showing the line of best fit.

We can introduce new input variable, t,representing years since 1994.

The least squares regression equation is:

$$C(t) = 113.318 + 2.209t \tag{7.3.12}$$

Using technology, the correlation coefficient was calculated to be 0.9965, suggesting a very strong increasing linear trend.

Using this to predict consumption in 2008 $(t = 14)$,

$$C(14) = 113.318 + 2.209(14) \tag{7.3.13}$$
$$= 144.244 \tag{7.3.14}$$

The model predicts 144.244 billion gallons of gasoline consumption in 2008.

Exercise 7.3.1

Use the model we created using technology in Example 7.3.6 to predict the gas consumption in 2011. Is this an interpolation or an extrapolation?

Answer available at the end of this section.

Answers

Exercise 7.3.1

150.871 billion gallons; extrapolation

This page titled 7.3: Fitting Linear Models to Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 2.4: Fitting Linear Models to Data by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

7.4: Modeling with Exponential Functions

Learning Objectives

Students will be able to:

- Make predictions by using the exponential growth (or decay) model
- Use the doubling-time and half-life models to make predictions
- Estimate the doubling time and half-life by using the Rule of 70

in order to apply mathematical modeling to solve real-world applications.

The next growth we will examine is exponential growth. Linear growth occurs by adding the same amount in each unit of time. Exponential growth happens when an initial population increases by the same percentage or factor over equal time increments or generations. This is known as relative growth and is usually expressed as a percentage. For example, let's say a population is growing by 5% each year. This means that 5% of the current population is added each year to the total. If the population was originally 10,000 in the first year, then 5% of 10,000 would be added which is 500, giving a new population of 10,500. In the second year, another 5% would be added, but this time you would take 5% of 10,500 which is 525, giving a new population of 11,025. While the percentage growth is the same each year, the actual number of people added changes, which makes this growth different from the linear model discussed earlier!

Definition: Exponential Growth

A quantity grows exponentially if it grows by a constant factor or rate for each unit of time.

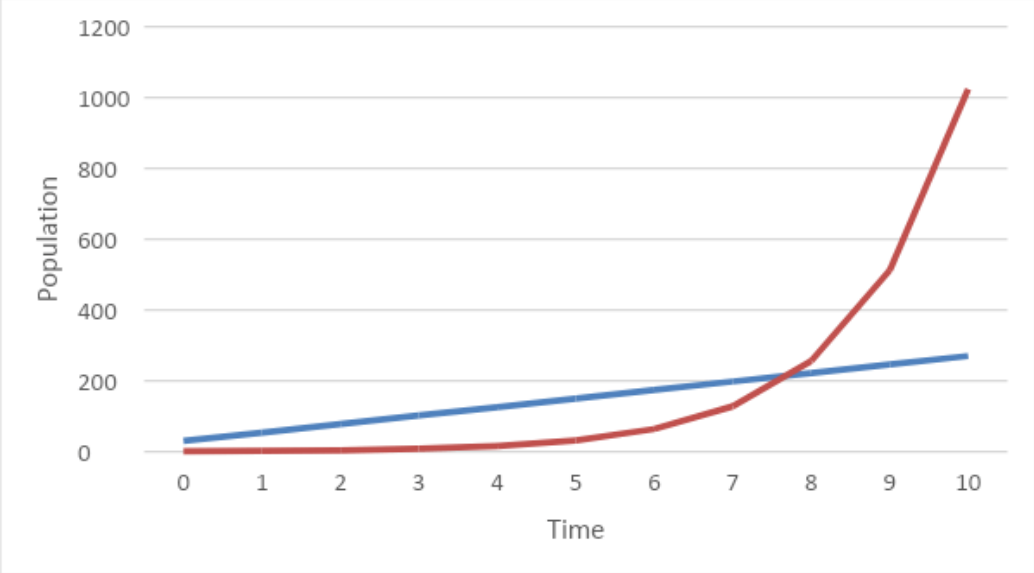


Figure 7.4.1: Graphical Comparison of Linear Growth (blue) and Exponential Growth (red)

In this graph, the blue straight line represents linear growth and the red curved line represents exponential growth.

Example 7.4.1: City Growth

A city is growing at a rate of 1.6% per year. The initial population in 2010 is $P_0 = 125,000$. Calculate the city's population over the next few years.

Solution

The relative growth rate is 1.6%. This means an additional 1.6% is added on to 100% of the population that already exists each year. This is a growth factor of 101.6%.

- Population in 2011

$$125,000(1.016)^1 = 127,000$$

- Population in 2012

$$127,000(1.016) = 125,000(1.016)^2 = 129,032$$

- Population in 2013

$$129,032(1.016) = 125,000(1.016)^3 = 131,097$$

We can create an equation for the city's growth. Each year the population is 101.6% more than the previous year.

$$P(t) = 125,000(1 + 0.016)^t$$



Figure 7.4.2: Graph of City Growth

Note: The graph of the city growth follows an exponential growth model

Example 7.4.2: A Shrinking Population

St. Louis, Missouri has declined in population at a rate of 1.6 % per year over the last 60 years. The population in 1950 was 857,000. Find the population in 2014. (Wikipedia, n.d.)

Solution

$P_0 = 857,000$

The relative rate of decrease (known as a decay rate) is 1.6%. This means 1.6% of the population is subtracted from 100% of the population that already exists each year. This is a factor of 98.4%.

- Population in 1951

$$857,000(0.984)^1 = 843,288$$

- Population in 1952

$$843,228(0.984) = 857,000(0.984)^2 = 829,795$$

- Population in 1953

$$828,795(0.984) = 857,000(0.984)^3 = 816,519$$

We can create an equation for the city's growth. Each year the population is 1.6% less than the previous year.

$$P(t) = 857,000(1 - 0.016)^t \tag{7.4.1}$$

So the population of St. Louis Missouri in 2014, when $t = 64$, is:

$$\begin{aligned} P(64) &= 857,000(1 - 0.016)^{64} \\ &= 857,000(0.984)^{64} \\ &= 305,258 \end{aligned}$$

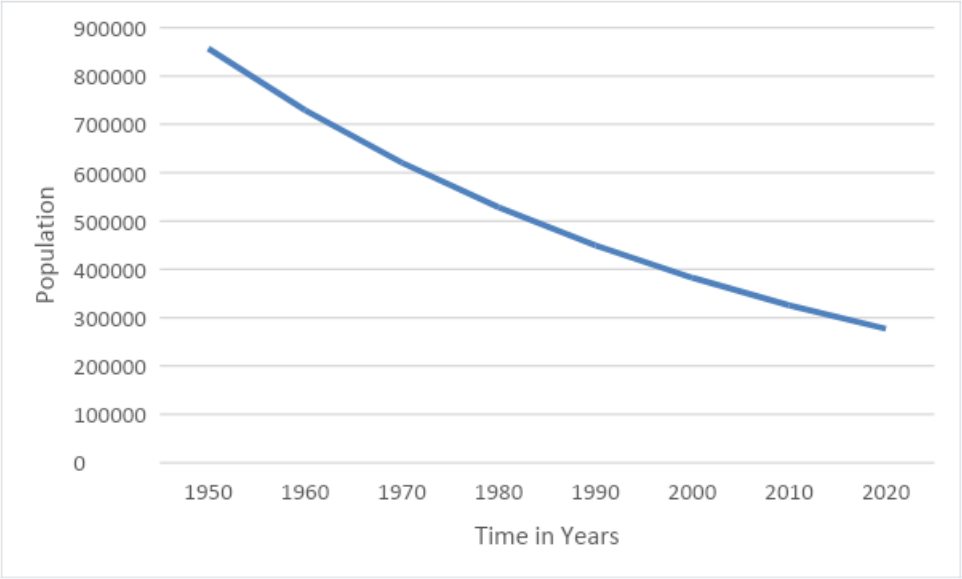


Figure 7.4.3: Graph of St. Louis, Missouri Population Decline

Note: The graph of the population of St. Louis, Missouri over time follows a declining exponential growth model.

Definition: Exponential Growth Model

$P(t) = P_0(1 + r)^t$

P_0 is the initial population.

r is the relative growth or decay rate.

$(1 + r)$ is the growth multiplier.

t is the time unit.

r is positive if the population is increasing (in cases of growth) and negative if the population is decreasing (in cases of decay).

Example 7.4.3: Inflation

The average inflation rate of the U.S. dollar over the last five years is 1.7% per year. If a new car cost $18,000 five years ago, how much would it cost today? (U.S. Inflation Calculator, n.d.)

Solution

To solve this problem, we use the exponential growth model with *r* = 1.7%.

$P_0 = 18,000$ and $t = 5$

$$P(t) = 18,000(1 + 0.017)^t$$
$$P(5) = 18,000(1 + 0.017)^5 = 19,582.91$$

This car would cost $\$19,582.91$ today.

Example 7.4.4: Ebola Epidemic in Sierra Leone

In May of 2014 there were 15 cases of Ebola in Sierra Leone. By August, there were 850 cases. If the virus is spreading at the same rate (exponential growth), how many cases will there be in February of 2015? (McKenna, 2014)

Solution

To solve this problem, we have to find three things; the growth rate per month, the exponential growth model, and the number of cases of Ebola in February 2015. First calculate the growth rate per month. To do this, use the initial population$P_0 = 15$, in May 2014. Also, in August, three months later, the number of cases was 850 so,$P(3) = 850$.

Use these values and the exponential growth model to solve for r.

$$\begin{aligned} P(t) &= P0(1 + r)^t \\ 850 &= 15(1 + r)^3 \\ 56.67 &= (1 + r)^3 \\ \sqrt[3]{56.67} &= \sqrt[3]{(1 + r)^3} \\ 3.84 &= 1 + r \\ 2.84 &= r \end{aligned}$$

The growth rate is 284% per month. Thus, the exponential growth model is:

$$P(t) = 15(1 + 2.84)^t = 15(3.84)^t$$

Now, we use this to calculate the number of cases of Ebola in Sierra Leone in February 2015, which is 9 months after the initial outbreak so, $t = 9$

$$P(9) = 15(3.84)^9 = 2,725,250 \tag{7.4.2}$$

If this same exponential growth rate continues, the number of Ebola cases in Sierra Leone in February 2015 would be 2,725,250.

This is a bleak prediction for the community of Sierra Leone. Fortunately, the growth rate of this deadly virus should be reduced by the world community and World Health Organization by providing the needed means to fight the initial spread.

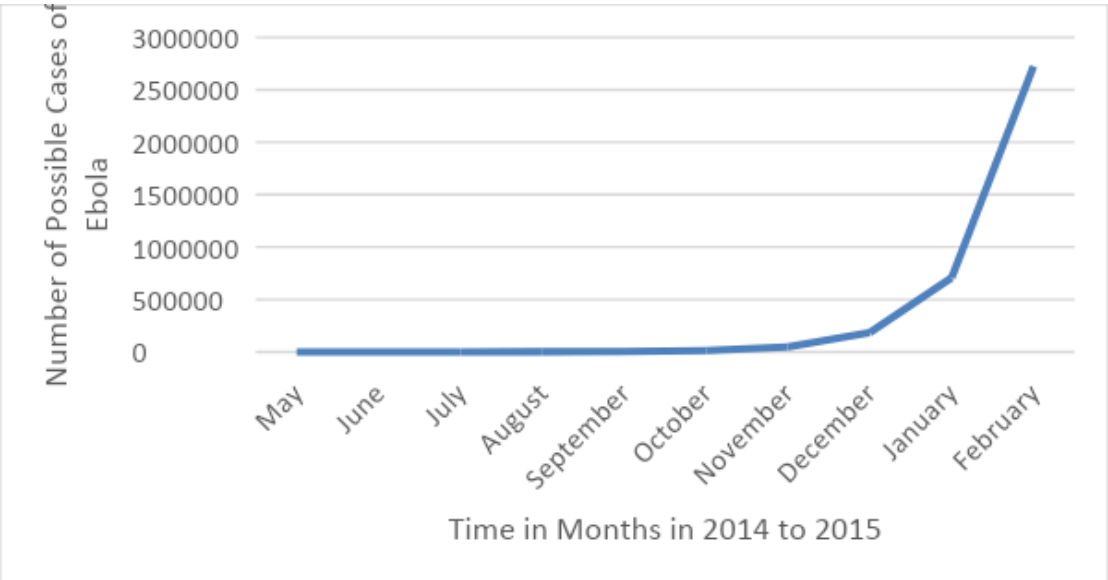


Figure 7.4.4: Graph of Ebola Virus is Sierra Leone

Note: The graph of the number of possible Ebola cases in Sierra Leone over time follows an increasing exponential growth model.

Example 7.4.5: Population Decline in Puerto Rico

According to a new forecast, the population of Puerto Rico is in decline. If the population in 2010 is 3,978,000 and the prediction for the population in 2050 is 3,697,000, find the annual percent decrease rate. (Bloomberg Businessweek, n.d.)

Solution

To solve this problem we use the exponential growth model. We need to solve for *r*.

$$\begin{aligned} P(t) &= P_0(1+r)^t \text{where } t = 40 \text{ years} \\ P(40) &= 3,697,000 \text{and} P_0 = 3,978,000 \\ 3,697,000 &= 3,978,000(1+r)^{40} \\ 0.92936 &= (1+r)^{40} \\ \sqrt[40]{0.92936} &= \sqrt[40]{(1+r)^{40}} \\ 0.99817 &= 1+r \\ -0.0018 &= r \end{aligned}$$

The annual percent decrease is 0.18%.

Example 7.4.6: April Fool's Joke

Let's say that on April 1st I say I will give you a penny, on April 2nd two pennies, four pennies on April 3rd, and that I will double the amount each day until the end of the month. How much money would I have agreed to give you on April 30? With $P_0 = \$0.01$, we get the following table:

Table 7.4.5: April Fool's Joke

Day	Dollar Amount
April 1 = P_0	0.01
April 2 = P_1	0.02
April 3 = P_2	0.04
April 4 = P_3	0.08
April 5 = P_4	0.16
April 6 = P_5	0.32
....	....
April 30 = P_{29}	?

In this example, the money received each day is 100% more than the previous day. If we use the exponential growth model $P(t) = P_0(1 + r)^t$ with *r* = 1, we get the doubling time model.

$$P(t) = P_0(1 + 1)^t = P_0(2)^t$$

We use it to find the dollar amount when $t = 29$ which represents April 30

$$P(29) = 0.01(2)^{29} = \$5,368,709.12$$

Surprised? That is a lot of pennies.

Doubling Time Model

Example 7.4.7: *E. coli* Bacteria

A water tank up on the San Francisco Peaks is contaminated with a colony of 80,000 *E. coli* bacteria. The population doubles every five days. We want to find a model for the population of bacteria present after t days. The amount of time it takes the population to double is five days, so this is our time unit. After t days have passed, then $t/5$ is the number of time units that have passed. Starting with the initial amount of 80,000 bacteria, our doubling model becomes:

$$P(t) = 80,000(2)^{\frac{t}{5}}$$

Using this model, how large is the colony in two weeks' time? We have to be careful that the units on the times are the same; 2 weeks = 14 days.

Solution

$$P(14) = 80,000(2)^{\frac{14}{5}} = 557,152$$

The colony is now 557,152 bacteria.

Definition: Doubling Time Model

If D is the doubling time of a quantity (the amount of time it takes the quantity to double) and P_0 is the initial amount of the quantity then the amount of the quantity present after t units of time is $P(t) = P_0(2)^{\frac{t}{D}}$

Example 7.4.8: Files

The doubling time of a population of flies is eight days. If there are initially 100 flies, how many flies will there be in 17 days?

Solution

To solve this problem, use the doubling time model with $D = 8$ and $P_0 = 100$ so the doubling time model for this problem is:

$$P(t) = 100(2)^{t/8}$$

When $t = 17\ days$,

$$P(17) = 100(2)^{\frac{17}{8}} = 436$$

There are 436 flies after 17 days.

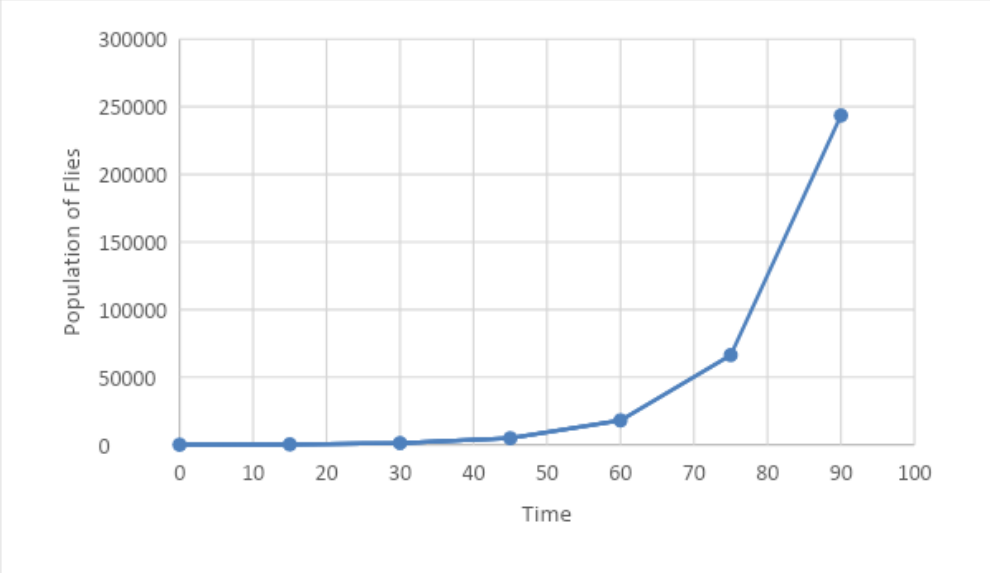


Figure 7.4.6: Graph of Fly Population in Three Months

Note: The population of flies follows an exponential growth model.

Example 7.4.9: Bacteria Growth

Suppose that a bacteria population doubles every six hours. If the initial population is 4000 individuals, how many hours would it take the population to increase to 25,000?

Solution

$P_0 = 4000$ and $D = 6$, so the doubling time model for this problem is:

$$P(t) = 4000(2)^{\frac{t}{6}}$$

Now, find t when $P(t) = 25,000$

$$25,000 = 4000(2)^{\frac{t}{6}}$$

$$\frac{25,000}{4000} = \frac{\cancel{4000}\,(2)^{\frac{t}{6}}}{\cancel{4000}}$$

$$6.25 = (2)^{\frac{t}{6}}$$

Now, we know that the exponent on the base of 2 has to be between 2 and 3 since

$$(2)^2 = 4$$

$$(2)^3 = 8$$

Since the exponent on the base of 2 is

$$\frac{t}{6} \tag{7.4.3}$$

this means t must be between 12 and 18 to ensure the exponent is between 2 and 3. By guessing different values of t, we can estimate the answer until we find a value of t that gives the proper results when substituted into the exponent. The best estimate for t is 16 since:

$$(2)^{\frac{t}{6}} = (2)^{\frac{16}{6}} = 6.3496$$

Which is very close to 6.25. Using technology would allow us to get a more accurate answer. Our conclusion is it would take approximately 16 hours for the population of the bacteria to increase to 25,000.

Example 7.4.10: Cancer Growth Rate

A certain cancerous tumor doubles in size every six months. If the initial size of the tumor is four cells, how many cells will there in three years? In seven years?

Solution

To calculate the number of cells in the tumor, we use the doubling time model. Change the time units to be the same. The doubling time is six months = 0.5 years.

$$P(t) = 4(2)^{\frac{t}{0.5}}$$

When $t = 3$ years

$$P(3) = 4(2)^{\frac{3}{0.5}} = 256\text{cells}$$

When $t = 7$ years

$$P(7) = 4(2)^{\frac{7}{0.5}} = 65,536\text{cells}$$

Rule of 70

There is a simple formula for approximating the doubling time of a population. It is called the rule of 70 and it is an approximation for growth rates less than 15%. Do not use this formula if the growth rate is 15% or greater.

Rule of 70

For a quantity growing at a constant percentage rate (not written as a decimal), R, per time period, the doubling time is approximately given by

$$D \approx \frac{70}{R} \tag{7.4.4}$$

Example 7.4.11: Bird Population

A bird population on a certain island has an annual growth rate of 2.5% per year. Approximate the number of years it will take the population to double. If the initial population is 20 birds, use it to find the bird population of the island in 17 years.

Solution

To solve this problem, first approximate the population doubling time.

Doubling time $D \approx \frac{70}{2.5} = 28$ years.

With the bird population doubling in 28 years, we use the doubling time model to find the population is 17 years.

$$P(t) = 20(2)^{\frac{t}{28}}$$

When $t = 17$ years

$$P(16) = 20(2)^{\frac{17}{28}} = 30.46$$

There will be 30 birds on the island in 17 years.

Example 7.4.12: Approximating Annual Growth Rate

Suppose that a certain city's population doubles every 12 years. What is the approximate annual growth rate of the city?

Solution

By solving the doubling time model for the growth rate, we can solve this problem.

$$\begin{aligned} D &\approx \frac{70}{R} \\ R \cdot D &\approx \frac{70}{\cancel{R}} \cdot \cancel{R} \\ RD &\approx 70 \\ \frac{R\cancel{D}}{\cancel{D}} &\approx \frac{70}{D} \\ \text{Annual growth rate R} &\approx \frac{70}{D} \\ R &= \frac{70}{12} = 5.83\% \end{aligned}$$

The annual growth rate of the city is approximately 5.83%

Exponential Decay and Half-Life Model

The half-life of a material is the time it takes for a quantity of material to be cut in half. This term is commonly used when describing radioactive metals like uranium or plutonium. For example, the half-life of carbon-14 is 5730 years, which means that 500 mg of carbon-14 will decay to 250 mg in 5370 years.

If a substance has a half-life, this means that half of the substance will be gone in a unit of time. In other words, the amount decreases by 50% per unit of time. Using the exponential growth model with a decrease of 50%, we have

$$P(t) = P_0(1 - 0.5)^t = P_0(\frac{1}{2})^t \tag{7.4.5}$$

Definition: Half-Life Model

If H is the half-life of a quantity (the amount of time it takes the quantity to be cut in half) and P_0 is the initial amount of the quantity then the amount of the quantity present after t units of time is

$$P(t) = P_0(\frac{1}{2})^{\frac{t}{H}} \tag{7.4.6}$$

Example 7.4.13: Half-Life

Let's say a substance has a half-life of eight days. If there are 40 grams present now, how much is left after three days?

Solution

We want to find a model for the quantity of the substance that remains after *t* days. The amount of time it takes the quantity to be reduced by half is eight days, so this is our time unit. After *t* days have passed, then t8 is the number of time units that have passed. Starting with the initial amount of 40, our half-life model becomes:

$$P(t) = 40(\frac{1}{2})^{\frac{t}{8}}$$

With $t = 3$

$$P(3) = 40(\frac{1}{2})^{\frac{3}{8}} = 30.8$$

There are 30.8 grams of the substance remaining after three days.

Example 7.4.14: Lead-209

Lead-209 is a radioactive isotope. It has a half-life of 3.3 hours. Suppose that 40 milligrams of this isotope is created in an experiment, how much is left after 14 hours?

Solution

Use the half-life model to solve this problem.

$P_0 = 40$ and $H = 3.3$, so the half-life model for this problem is:

$$P(t) = 40(\frac{1}{2})^{\frac{t}{3.3}}$$

With $t = 14$ hours,

$$P(14) = 40(\frac{1}{2})^{\frac{14}{3.3}} = 2.1$$

There are 2.1 milligrams of Lead-209 remaining after 14 hours.

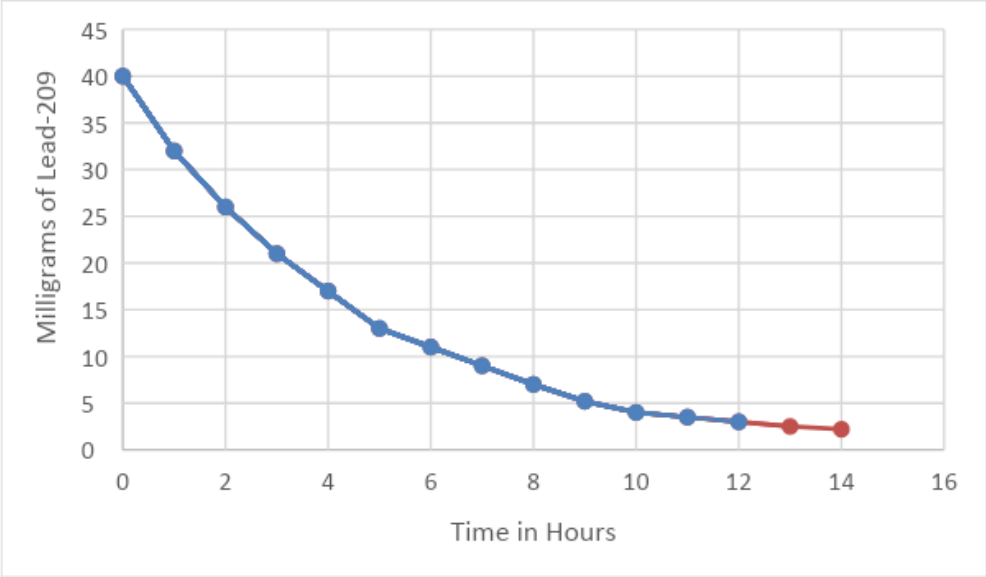


Figure 7.4.7: Lead-209 Decay Graph

Note: The milligrams of Lead-209 remaining follows a decreasing exponential growth model.

Example 7.4.15: Nobelium-259

Nobelium-259 has a half-life of 58 minutes. If you have 1000 grams, how much will be left in two hours?

Solution

We solve this problem using the half-life model. Before we begin, it is important to note the time units. The half-life is given in minutes and we want to know how much is left in two hours. Convert hours to minutes when using the model: two hours = 120 minutes.

$P_0 = 1000$ and $H = 58$ minutes, so the half-life model for this problem is:

$$P(t) = 1000(\frac{1}{2})^{\frac{t}{58}}$$

With $t = 120$ minutes,

$$P(120) = 1000(\frac{1}{2})^{\frac{120}{58}} = 238.33$$

There are 238 grams of Nobelium-259 is remaining after two hours.

Example 7.4.16: Carbon-14

Radioactive carbon-14 is used to determine the age of artifacts because it concentrates in organisms only when they are alive. It has a half-life of 5730 years. In 1947, earthenware jars containing what are known as the Dead Sea Scrolls were found. Analysis indicated that the scroll wrappings contained 76% of their original carbon-14. Estimate the age of the Dead Sea Scrolls.

Solution

In this problem, we want to estimate the age of the scrolls. In 1947, 76% of the carbon-14 remained. This means that the amount remaining at time t divided by the original amount of carbon-14, P_0, is equal to 76%. So, $\frac{P(t)}{P_0} = 0.76$ we use this fact to solve for *t*.

$$P(t) = P_0(\frac{1}{2})^{\frac{t}{5730}}$$
$$\frac{P(t)}{P_0} = (\frac{1}{2})^{\frac{t}{5730}}$$
$$0.76 = (\frac{1}{2})^{\frac{t}{5730}}$$

Now, we know that the exponent on the base of 1/2 has to be between 0 and 1 since

$$(\frac{1}{2})^0 = 1$$

$$(\frac{1}{2})^1 = 0.5$$

Since the exponent on the base of 1/2 is

$$\frac{t}{5730} \tag{7.4.7}$$

this means t must be between 0 and 5730 to ensure the exponent is between 0 and 1. By guessing different values of t, we can estimate the answer until we find a value of t that gives the proper results when substituted into the exponent. The best estimate for t is 2270 since:

$$(\frac{1}{2})^{\frac{t}{5730}} = (\frac{1}{2})^{\frac{2270}{5730}} = 0.76$$

Which is almost exactly what we needed. Using technology would allow us to get a more accurate answer and more importantly, it would make it much easier to find the correct value since there are many choices for t between 0 and 5730.

Our conclusion is the Dead Sea Scrolls are approximately 2270 years old.

Rule of 70 for Half-Life:

There is simple formula for approximating the half-life of a population. It is called the rule of 70 and is an approximation for decay rates less than 15%. Do not use this formula if the decay rate is 15% or greater.

Rule of 70

For a quantity decreasing at a constant percentage (not written as a decimal), *R*, per time period, the half-life is approximately given by:

$$\text{Half-life } H \approx \frac{70}{R} \tag{7.4.8}$$

Example 7.4.17: Elephant Population

The population of wild elephants is decreasing by 7% per year. Approximate the half -life for this population. If there are currently 8000 elephants left in the wild, how many will remain in 25 years?

Solution

To solve this problem, use the half-life approximation formula.

$$\text{Half-Life } H \approx \frac{70}{7} = 10 \text{ years}$$

$P_0 = 7000$ and $H = 10$ years, so the half-life model for this problem is:

$$P(t) = 7000(\frac{1}{2})^{\frac{t}{10}}$$

When $t = 25$,

$$P(25) = 7000(\frac{1}{2})^{\frac{25}{10}} = 1237.4$$

There will be approximately 1237 wild elephants left in 25 years.

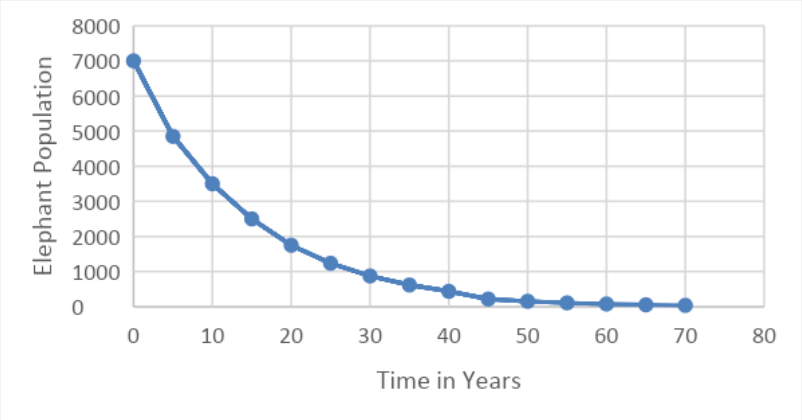


Figure 7.4.8: Elephant Population over a 70 Year Span.

Note: The population of elephants follows a decreasing exponential growth model.

This page titled 7.4: Modeling with Exponential Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier via source content that was edited to the style and standards of the LibreTexts platform.

- 4.2: Exponential Growth by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.
- 4.3: Special Cases- Doubling Time and Half-Life by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.

7.5: Fitting Exponential Models to Data

Learning Objectives

- Use a graphing utility to find the exponential model that best fits the data
- Make predictions using an exponential model found with technology
- Determine if a linear or an exponential model fits the data better

in order to apply mathematical modeling to solve real-world applications.

In previous sections of this chapter, we were either given a function explicitly to graph or evaluate, or we were given a set of points that were guaranteed to lie on the curve. Then we used algebra to find the equation that fit the points exactly. In this section, we use a modeling technique called *regression analysis* to find a curve that models data collected from real-world observations. With regression analysis, we don't expect all the points to lie perfectly on the curve. The idea is to find a model that best fits the data. Then we use the model to make predictions about future events.

Do not be confused by the word *model*. In mathematics, we often use the terms *function*, *equation*, and *model* interchangeably, even though they each have their own formal definition. The term *model* is typically used to indicate that the equation or function approximates a real-world situation.

We will concentrate on three types of regression models in this section: exponential, logarithmic, and logistic. Having already worked with each of these functions gives us an advantage. Knowing their formal definitions, the behavior of their graphs, and some of their real-world applications gives us the opportunity to deepen our understanding. As each regression model is presented, key features and definitions of its associated function are included for review. Take a moment to rethink each of these functions, reflect on the work we've done so far, and then explore the ways regression is used to model real-world phenomena.

Building an Exponential Model from Data

As we've learned, there are a multitude of situations that can be modeled by exponential functions, such as investment growth, radioactive decay, atmospheric pressure changes, and temperatures of a cooling object. What do these phenomena have in common? For one thing, all the models either increase or decrease as time moves forward. But that's not the whole story. It's the *way* data increase or decrease that helps us determine whether it is best modeled by an exponential equation. Knowing the behavior of exponential functions in general allows us to recognize when to use exponential regression, so let's review exponential growth and decay.

Recall that exponential functions have the form $y = ab^x$. When performing regression analysis, we use the form most commonly used on graphing utilities, $y = ab^x$. Take a moment to reflect on the characteristics we've already learned about the exponential function $y = ab^x$ (assume $a > 0$):

- b must be greater than zero and not equal to one.
- The initial value of the model is $y = a$.
  - If $b > 1$, the function models exponential growth. As x increases, the outputs of the model increase slowly at first, but then increase more and more rapidly, without bound.
  - If $0 < b < 1$, the function models **exponential decay**. As x increases, the outputs for the model decrease rapidly at first and then level off to become asymptotic to the *x*-axis. In other words, the outputs never become equal to or less than zero.

As part of the results, your calculator will display a number known as the *correlation coefficient*, labeled by the variable r, or r^2. (You may have to change the calculator's settings for these to be shown.) The values are an indication of the "goodness of fit" of the regression equation to the data. We more commonly use the value of r^2 instead of r, but the closer either value is to 1, the better the regression equation approximates the data.

Exponential Regression

Exponential regression is used to model situations in which growth begins slowly and then accelerates rapidly without bound, or where decay begins rapidly and then slows down to get closer and closer to zero. We use the command “ExpReg” on a graphing utility to fit an exponential function to a set of data points. We can also use desmos instead to find an exponential function that fits the data. This returns an equation of the form,

$$y = ab^x \tag{7.5.1}$$

Note that:

- b must be non-negative.
- when $b > 1$, we have an exponential growth model.
- when $0 < b < 1$, we have an exponential decay model.

How to: Given a set of data, perform exponential regression using desmos

When using desmos, you will first, create a table and fill in the two columns with the data where the first column is x_1 and the second column is y_1. Then, type "$y_1 \sim a\, b\hat{}x_1$" and desmos will create the best fit exponential function and also give the values of a and b. Please make sure to check the box called "Log Mode" if you want demos to output the same answers as other graphing utilities.

Example 7.5.1: Using Exponential Regression to Fit a Model to Data

In 2007, a university study was published investigating the crash risk of alcohol impaired driving. Data from $2,871$ crashes were used to measure the association of a person’s blood alcohol level (BAC) with the risk of being in an accident. Table $7.5.1$ shows results from the study. The *relative risk* is a measure of how many times more likely a person is to crash. So, for example, a person with a BAC of 0.09 is 3.54 times as likely to crash as a person who has not been drinking alcohol.

Table $7.5.1$

BAC	0	0.01	0.03	0.05	0.07	0.09
Relative Risk of Crashing	1	1.03	1.06	1.38	2.09	3.54
BAC	0.11	0.13	0.15	0.17	0.19	0.21
Relative Risk of Crashing	6.41	12.6	22.1	39.05	65.32	99.78

a. Let x represent the BAC level, and let y represent the corresponding relative risk. Use exponential regression to fit a model to these data.
b. After 6 drinks, a person weighing 160 pounds will have a BAC of about 0.16. How many times more likely is a person with this weight to crash if they drive after having a 6-pack of beer? Round to the nearest hundredth.

Solution

a. Using the STAT then EDIT menu on a graphing utility, list the BAC values in L1 and the relative risk values in L2. Then use the STATPLOT feature to verify that the scatterplot follows the exponential pattern shown in Figure $7.5.1$:

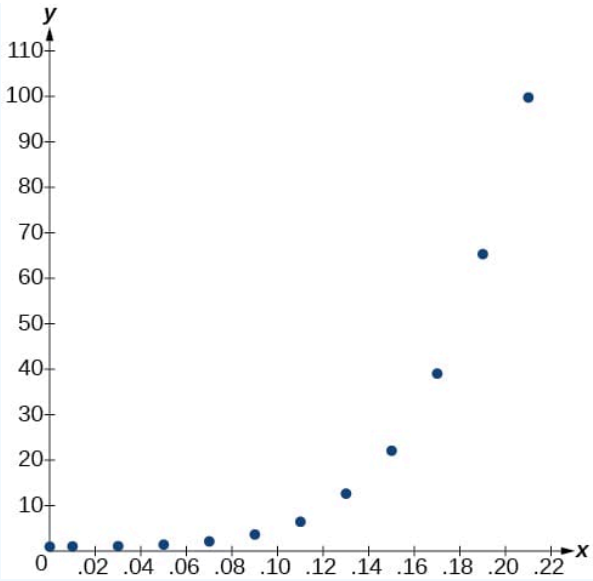


Figure $7.5.1$

Use the "ExpReg" command from the STAT then CALC menu to obtain the exponential model,

$$y = 0.58304829(2.20720213E10)^x$$

Converting from scientific notation, we have:

$$y = 0.58304829(22072021300)^x$$

Notice that $r^2 \approx 0.97$ which indicates the model is a good fit to the data. To see this, graph the model in the same window as the scatterplot to verify it is a good fit as shown in Figure $7.5.2$:

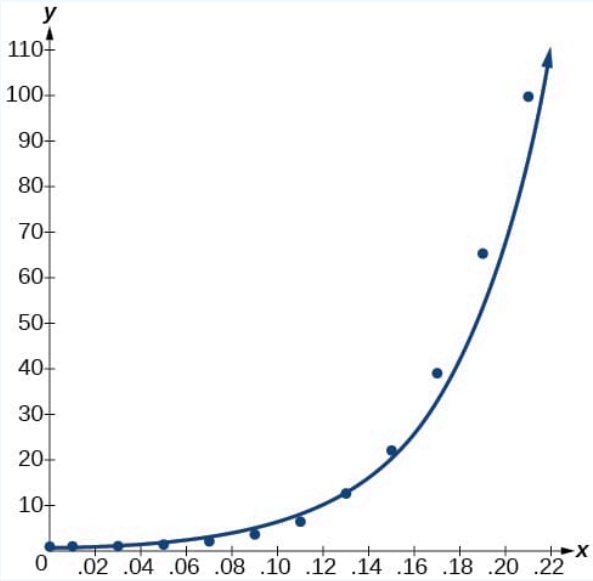


Figure $7.5.2$

b. Use the model to estimate the risk associated with a BAC of 0.16. Substitute 0.16 for x in the model and solve for y.

$$
\begin{aligned}
y &= 0.58304829(22072021300)^{x} && \text{Use the regression model found in part } (a)\\
&= 0.58304829(22072021300)^{0.16} && \text{Substitute } 0.16 \text{ for x}\\
&\approx 26.35 && \text{Round to the nearest hundredth}
\end{aligned}
$$

If a 160-pound person drives after having 6 drinks, he or she is about 26.35 times more likely to crash than if driving while sober.

Here is the same problem done using desmos. If "Log Mode" is not checked, desmos will provide a more accurate model that does not match the results of many other graphing utilities including those used to obtain the results in this example. You can view the problem in desmos by clicking https://www.desmos.com/calculator/ev2papvkzo.

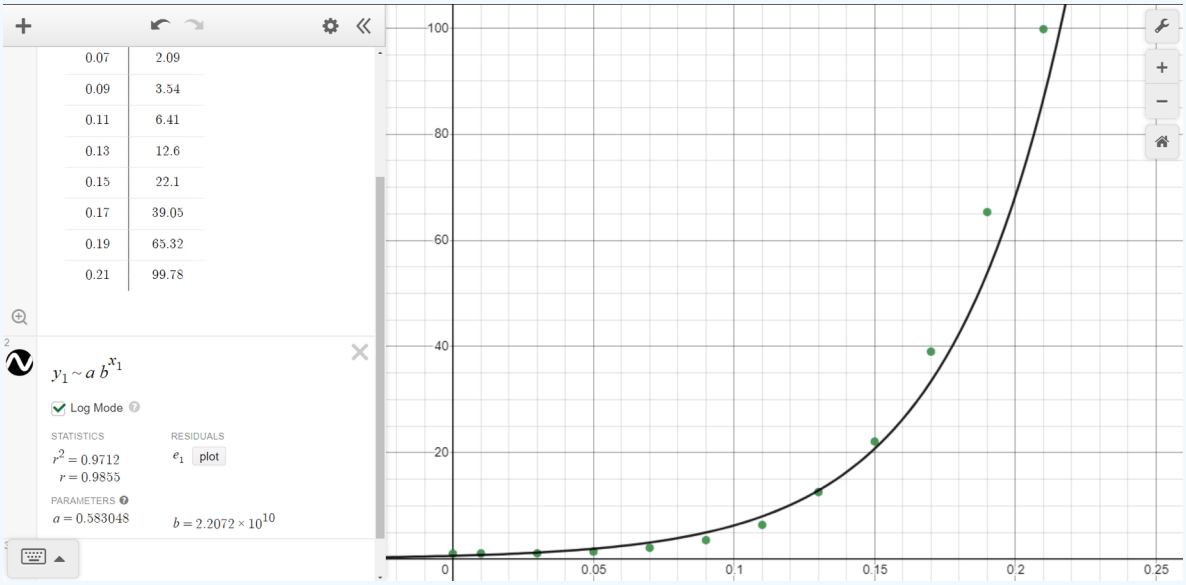


Figure $7.5.3$

Exercise $7.5.1$

Table $7.5.2$ shows a recent graduate's credit card balance each month after graduation.

Table $7.5.2$

Month	1	2	3	4	5	6	7	8
Debt ($)	620.00	761.88	899.80	1039.93	1270.63	1589.04	1851.31	2154.92

a. Use exponential regression to fit a model to these data.
b. If spending continues at this rate, what will the graduate's credit card debt be one year after graduating?

Answer a available at the end of this section.

Answer b available at the end of this section.

Q&A: Is it reasonable to assume that an exponential regression model will represent a situation indefinitely?

No. Remember that models are formed by real-world data gathered for regression. It is usually reasonable to make estimates within the interval of original observation (interpolation). However, when a model is used to make predictions, it is important to use reasoning skills to determine whether the model makes sense for inputs far beyond the original observation interval (extrapolation).

Answers

Exercise $7.5.1$

The exponential regression model that fits these data is $y = 522.88585984(1.19645256)^x$.

Exercise $7.5.1$

If spending continues at this rate, the graduate's credit card debt will be $\$4,499.38$ after one year.

This page titled 7.5: Fitting Exponential Models to Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 3.2: Quadratic Functions by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.
- 4.2: Exponential Growth by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.

7.6: Putting It All Together

Learning Objectives

Students will be able to:

- Apply both a linear and an exponential regression to a single set of data
- Determine if a linear or exponential model is the better fit for the data
- Construct and use a model to make predictions based on the data

in order to apply mathematical modeling to solve real-world applications.

Using Models to Make Predictions

So far, we have explored both linear and exponential models for a given set of data. Our goal is to now take a given set of data and decide which is the better model to use, and ultimately to use the model to make a prediction of an output for a given input.

As a reminder, a scatter plot can be used to show the relationship (or correlation) between two related values: the input and the output variables. In this section, we shall create a scatter plot for a given set of data, and then use technology to determine a model that will allow us to make predictions.

One of the objectives in creating a model is to **make predictions** about values we have not yet measured by finding the mathematical relationship between the input and the output variables, and relate this to our theoretical understanding. In so doing we derive a mathematical relationship where y is a function of x.

$$\begin{gathered} y = \text{function of x} \\ or \\ y = f(x) \end{gathered} \tag{7.6.1}$$

In this statement y is the output variable and plotted on the vertical axis and x is the input variable and plotted on the horizontal axis. Empirical data is collected through experimentation where x is changed and y is measured, with each data point being represented on a graph with the values (x,y). We have looked at linear and exponential functions so far, and the end of this chapter also discusses quadratic functions as another possible model.

$$y = mx + b \qquad (\text{linear function}) \tag{7.6.2}$$
$$y = ab^x \qquad (\text{exponential function}) \tag{7.6.3}$$
$$y = ax^2 + bx + c \qquad (\text{quadratic function}) \tag{7.6.4}$$

Each of the above equations has two variables (x,y) and two constants (m,b) or (a,b) except for the quadratic function which has three constants (a,b,c) and you should already know how to calculate the constants m(slope) and b (y-intercept) for a straight line from earlier in this chapter.

Our goals are three fold:

1. Use desmos to create a scatter plot of the given data.
2. Apply a linear and an exponential model and see which fits the data better.
3. Make a prediction of the value of an output by substituting the input into the better model.

In order to determine the better model, we need to pay attention to the value of r^2 that desmos gives us, which is the square of the correlation coefficient (r) mentioned in precious sections. When r^2=1 the equation perfectly predicts the output variable and the data follows the equation exactly. When r^2=0, the equation provides no correlation between the input and output variable. The closer r^2 is to 1, the more likely the model will provide accurate predictions, but it's also important to also look at the graph of the models and see if they reasonably fit the scatter plots. **Do not blindly make predictions based on the value of r^2.** You must always look at the phenomena you are studying to check for any predictions that are not physically possible. For example, if your

input variable is the size of a rock, and your output variable is the weight of a rock, any model that predicts a negative weight for a given size wouldn't be reasonable.

Lets look at some data and fit it to the two different models we have explored in this chapter.

Example 7.6.1

Table 7.6.1 shows a recent graduate's credit card balance each month after graduation.

Table 7.6.1

Month	1	2	3	4	5	6	7	8
Debt ($)	620.00	761.88	899.80	1039.93	1270.63	1589.04	1851.31	2154.92

a. Use a linear regression to fit a model to these data.

We will use desmos to recreate the table which will automatically generate a scatter plot. Then, following the same steps from the section on linear models, we type "$y_1 \sim mx_1 + b$" and desmos gives us "m = 220.119" and "b = 282.902". We are also given $r^2 = 0.9721$. This means the complete model is:

$y = 220.11\,x + 282.902$

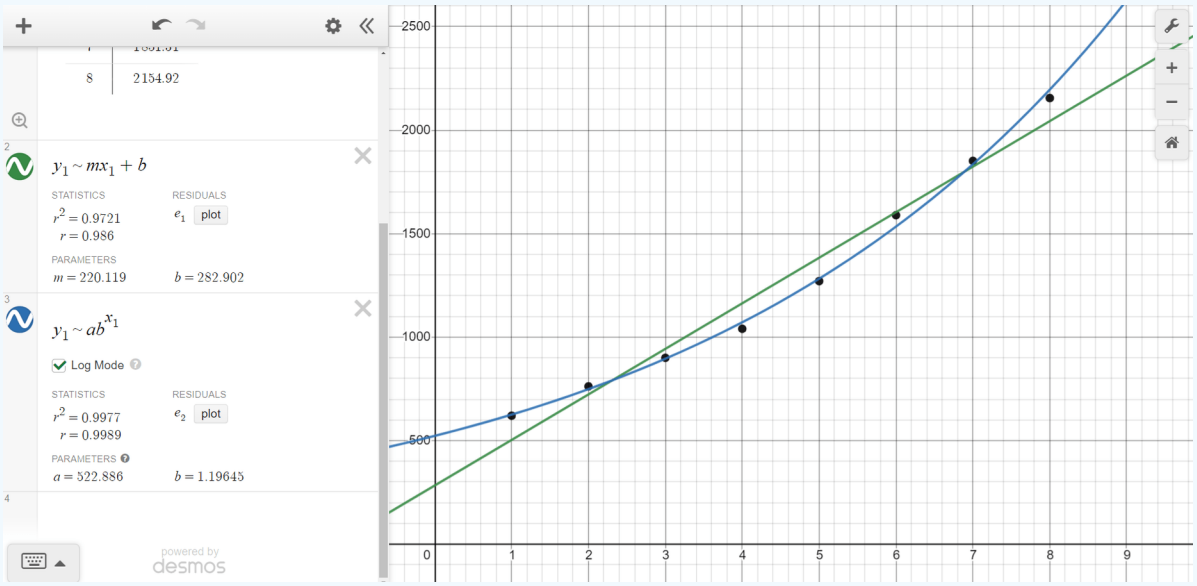
b. Use an exponential regression to fit a model to these data.

Using the same table already created in desmos from the previous step, we add the exponential model by typing "$y_1 \sim ab^{x_1}$" and desmos gives us "a = 522.886" and "b = 1.19645". We are also given $r^2 = 0.9977$. Please remember to click the box "Log Mode" to ensure you get the same results. This means the complete model is:

$y = 522.886(1.19645^x)$

c. Which model appears to fit the data better?

Looking at the values of r^2, it seems the exponential model is better. This is confirmed by visually looking at the models compared to the actual scatter plot. The figure below is what you should see on desmos.



d. Using the better model, make a prediction of the graduate's credit card debt one year after graduating.

We will use the exponential model to make this prediction. Make sure to plug in 12 months and not 1 year since the original data is given in terms of months! We get the following:

$y = 522.886(1.19645^{x}) = 522.886(1.19645^{12}) = 4499.38$

The graduate's credit card debt will be approximately $4,499.38 after one year. Had you used the linear model, you would have calculated a much different prediction of $2,924.33.

You can find this problem on desmos by clicking https://www.desmos.com/calculator/cx8igorint.

This page titled 7.6: Putting It All Together is shared under a license and was authored, remixed, and/or curated by Robert Belford.

- 4.2: Exponential Growth by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, and Kim Sonier is licensed CC BY-SA 4.0. Original source: https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier.

7.7: Modeling with Quadratic Functions

Learning Objectives

- Recognize characteristics of parabolas
- Understand how the graph of a parabola is related to its quadratic function
- Solve problems involving a quadratic function's minimum or maximum value

in order to apply mathematical modeling to solve real-world applications.

Curved antennas, such as the ones shown in Figure 7.7.1, are commonly used to focus microwaves and radio waves to transmit television and telephone signals, as well as satellite and spacecraft communication. The cross-section of the antenna is in the shape of a parabola, which can be described by a quadratic function.

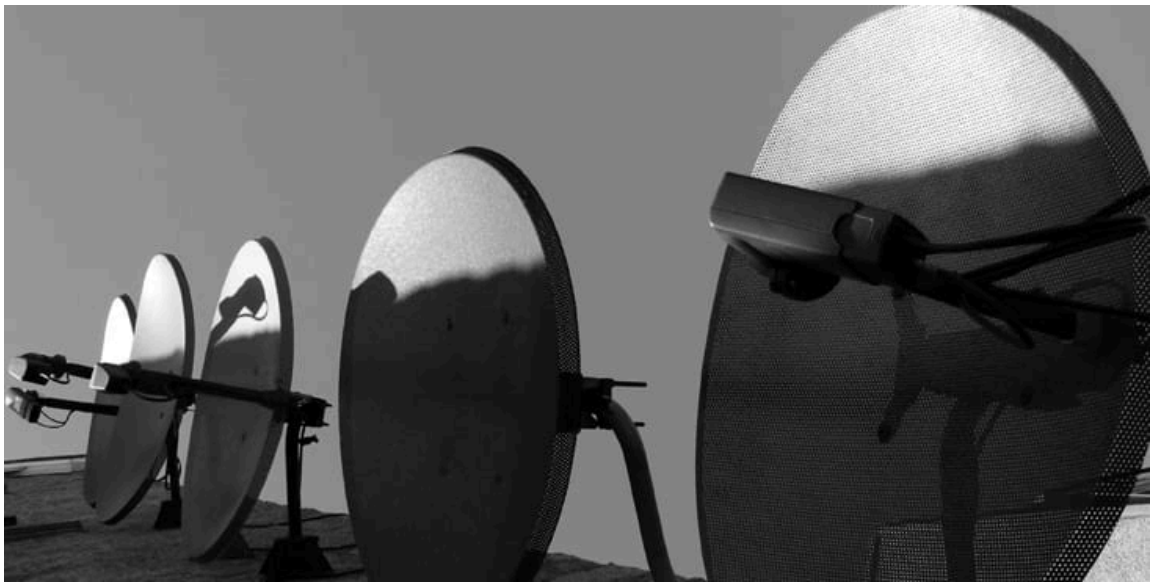


***Figure** 7.7.1**:** An array of satellite dishes. (credit: Matthew Colvin de Valle, Flickr)*

In this section, we will investigate quadratic functions, which frequently model problems involving area and projectile motion. Working with quadratic functions can be less complex than working with higher degree functions, so they provide a good opportunity for a detailed study of function behavior.

Recognizing Characteristics of Parabolas

The graph of a quadratic function is a U-shaped curve called a parabola. One important feature of the graph is that it has an extreme point, called the **vertex**. If the parabola opens up, the vertex represents the lowest point on the graph, or the **minimum value** of the quadratic function. If the parabola opens down, the vertex represents the highest point on the graph, or the **maximum value**. In either case, the vertex is a turning point on the graph. The graph is also symmetric with a vertical line drawn through the vertex, called the **axis of symmetry**. These features are illustrated in Figure 7.7.2.

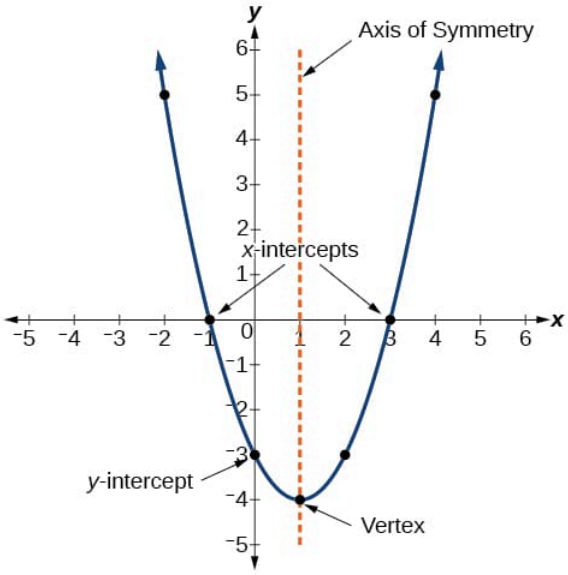


Figure 7.7.2: Graph of a parabola showing where the x and y intercepts, vertex, and axis of symmetry are.

The y-intercept is the point at which the parabola crosses the y-axis. The x-intercepts are the points at which the parabola crosses the x-axis. If they exist, the x-intercepts represent the **zeros**, or **roots**, of the quadratic function, the values of x at which $y = 0$.

Example 7.7.1: Identifying the Characteristics of a Parabola

Determine the vertex, axis of symmetry, zeros, and y-intercept of the parabola shown in Figure 7.7.3.

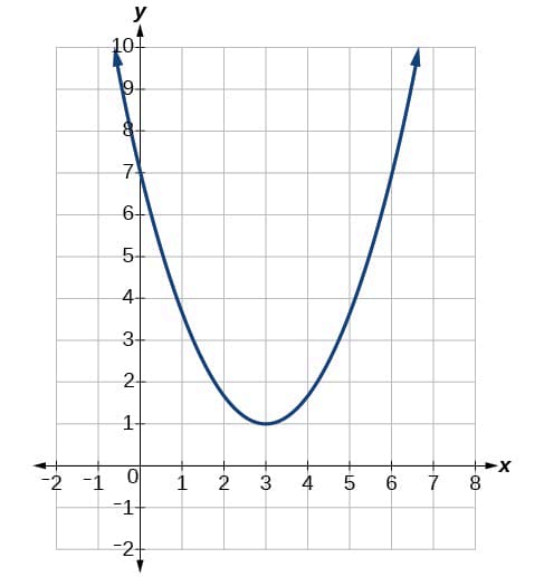


Figure 7.7.3.

Solution

The vertex is the turning point of the graph. We can see that the vertex is at $(3, 1)$. Because this parabola opens upward, the axis of symmetry is the vertical line that intersects the parabola at the vertex. So the axis of symmetry is $x = 3$. This parabola does not cross the x-axis, so it has no zeros. It crosses the y-axis at $(0, 7)$ so this is the y-intercept.

Understanding How the Graphs of Parabolas are Related to Their Quadratic Functions

The **general form of a quadratic function** presents the function in the form

$$f(x) = ax^2 + bx + c \tag{7.7.1}$$

where a, b, and c are real numbers and $a \neq 0$. If $a > 0$, the parabola opens upward. If $a < 0$, the parabola opens downward. We can use the general form of a parabola to find the equation for the axis of symmetry.

The axis of symmetry is defined by $x = -\frac{b}{2a}$. If we use the quadratic formula, $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, to solve $ax^2 + bx + c = 0$ for the x-intercepts, or zeros, we find the value of x halfway between them is always $x = -\frac{b}{2a}$, the equation for the axis of symmetry.

Figure $7.7.4$ represents the graph of the quadratic function written in general form as $y = x^2 + 4x + 3$. In this form, $a = 1$, $b = 4$, and $c = 3$. Because $a > 0$, the parabola opens upward. The axis of symmetry is $x = -\frac{4}{2(1)} = -2$. This also makes sense because we can see from the graph that the vertical line $x = -2$ divides the graph in half. The vertex always occurs along the axis of symmetry. For a parabola that opens upward, the vertex occurs at the lowest point on the graph, in this instance, $(-2, -1)$. The x-intercepts, those points where the parabola crosses the x-axis, occur at $(-3, 0)$ and $(-1, 0)$.

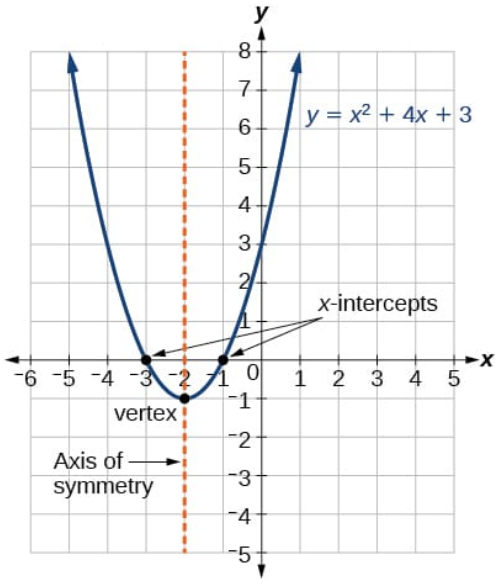


Figure $7.7.4$: Graph of a parabola showing where the x and y intercepts, vertex, and axis of symmetry are for the function $y = x^2 + 4x + 3$.

The **standard form of a quadratic function** presents the function in the form

$$f(x) = a(x - h)^2 + k \tag{7.7.2}$$

where (h, k) is the vertex. Because the vertex appears in the standard form of the quadratic function, this form is also known as the **vertex form of a quadratic function**.

As with the general form, if $a > 0$, the parabola opens upward and the vertex is a minimum. If $a < 0$, the parabola opens downward, and the vertex is a maximum. Figure $7.7.5$ represents the graph of the quadratic function written in standard form as $y = -3(x + 2)^2 + 4$. Since $x–h = x + 2$ in this example, $h = –2$. In this form, $a = -3$, $h = -2$, and $k = 4$. Because $a < 0$, the parabola opens downward. The vertex is at $(-2, 4)$.

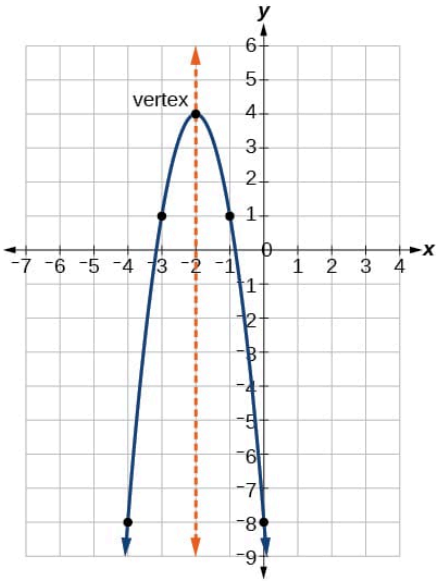


Figure 7.7.5: Graph of a parabola showing where the x and y intercepts, vertex, and axis of symmetry are for the function $y = -3(x + 2)^2 + 4$.

The standard form is useful for determining how the graph is transformed from the graph of $y = x^2$. Figure 7.7.6 is the graph of this basic function.

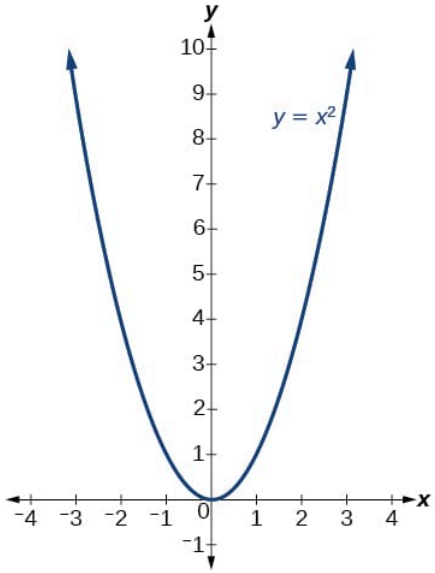


Figure 7.7.6: Graph of $y = x^2$.

If $k > 0$, the graph shifts upward, whereas if $k < 0$, the graph shifts downward. In Figure 7.7.5, $k > 0$, so the graph is shifted 4 units upward. If $h > 0$, the graph shifts toward the right and if $h < 0$, the graph shifts to the left. In Figure 7.7.5, $h < 0$, so the graph is shifted 2 units to the left. The magnitude of a indicates the stretch of the graph. If $|a| > 1$, the point associated with a particular x-value shifts farther from the x-axis, so the graph appears to become narrower, and there is a vertical stretch. But if $|a| < 1$, the point associated with a particular x-value shifts closer to the x-axis, so the graph appears to become wider, but in fact there is a vertical compression. In Figure 7.7.5, $|a| > 1$, so the graph becomes narrower.

The standard form and the general form are equivalent methods of describing the same function. We can see this by expanding out the general form and setting it equal to the standard form.

$$a(x - h)^2 + k = ax^2 + bx + c$$
$$ax^2 - 2ahx + (ah^2 + k) = ax^2 + bx + c$$

For the linear terms to be equal, the coefficients must be equal.

$$-2ah = b, \text{ so } h = -\frac{b}{2a}.$$

This is the **axis of symmetry** we defined earlier. Setting the constant terms equal:

$$\begin{aligned} ah^2 + k &= c \\ k &= c - ah^2 \\ &= c - a - \left(\frac{b}{2a}\right)^2 \\ &= c - \frac{b^2}{4a} \end{aligned}$$

In practice, though, it is usually easier to remember that k is the output value of the function when the input is h, so $f(h) = k$.

Definitions: Forms of Quadratic Functions

A quadratic function is a function of degree two. The graph of a *quadratic function* is a parabola.

- The *general form* of a quadratic function is $f(x) = ax^2 + bx + c$ where a, b, and c are real numbers and $a \neq 0$.
- The standard form of a quadratic function is $f(x) = a(x - h)^2 + k$.
- The vertex (h, k) is located at

$$h = -\frac{b}{2a}, \; k = f(h) = f(\frac{-b}{2a}). \tag{7.7.3}$$

HOWTO: Write a quadratic function in a general form

Given a graph of a quadratic function, write the equation of the function in general form.

1. Identify the horizontal shift of the parabola; this value is h. Identify the vertical shift of the parabola; this value is k.
2. Substitute the values of the horizontal and vertical shift for h and k. in the function $f(x) = a(x-h)^2 + k$.
3. Substitute the values of any point, other than the vertex, on the graph of the parabola for x and $f(x)$.
4. Solve for the stretch factor, $|a|$.
5. If the parabola opens up, $a > 0$. If the parabola opens down, $a < 0$ since this means the graph was reflected about the x-axis.
6. Expand and simplify to write in general form.

Example 7.7.2: Writing the Equation of a Quadratic Function from the Graph

Write an equation for the quadratic function g in Figure 7.7.7 as a transformation of $f(x) = x^2$, and then expand the formula, and simplify terms to write the equation in general form.

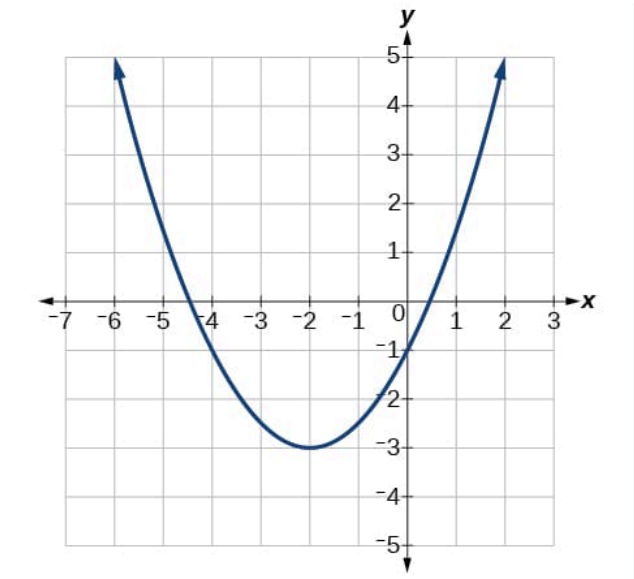


Figure 7.7.7: Graph of a parabola with its vertex at $(-2,-3)$.

Solution

We can see the graph of g is the graph of $f(x)=x^2$ shifted to the left 2 and down 3, giving a formula in the form $g(x)=a(x+2)^2$–3.

Substituting the coordinates of a point on the curve, such as $(0,-1)$, we can solve for the stretch factor.

$$-1=a(0+2)^2-3 \tag{7.7.4}$$
$$2=4a \tag{7.7.5}$$
$$a=\frac{1}{2} \tag{7.7.6}$$

In standard form, the algebraic model for this graph is $g(x)=\frac{1}{2}(x+2)^2$–3.

To write this in general polynomial form, we can expand the formula and simplify terms.

$$g(x)=\frac{1}{2}(x+2)^2-3 \tag{7.7.7}$$
$$=\frac{1}{2}(x+2)(x+2)-3 \tag{7.7.8}$$
$$=\frac{1}{2}(x^2+4x+4)-3 \tag{7.7.9}$$
$$=\frac{1}{2}x^2+2x+2-3 \tag{7.7.10}$$
$$=\frac{1}{2}x^2+2x-1 \tag{7.7.11}$$

Notice that the horizontal and vertical shifts of the basic graph of the quadratic function determine the location of the vertex of the parabola; the vertex is unaffected by stretches and compressions.

Analysis

We can check our work using the table feature on a graphing utility. First enter $\mathrm{Y1}=\frac{1}{2}(\mathrm{x}+2)^2-3$. Next, select TBLSET, then use $\mathrm{TblStart}=-6$ and $\Delta\mathrm{Tbl}=2$, and select TABLE. See Table 7.7.1

Table 7.7.1

x	-6	-4	-2	0	2

y	-5	-1	-3	-1	5

The ordered pairs in the table correspond to points on the graph.

Exercise 7.7.2

A coordinate grid has been superimposed over the quadratic path of a basketball in Figure 7.7.8. Find an equation for the path of the ball. Does the shooter make the basket?

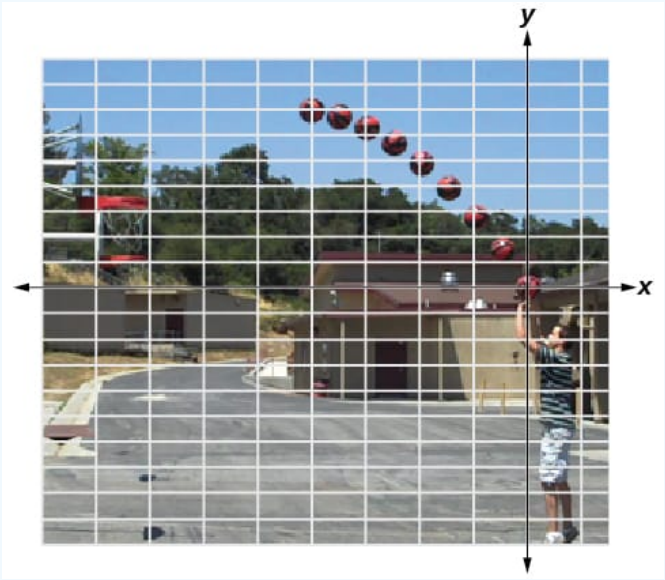


Figure 7.7.8***:*** *Stop motioned picture of a boy throwing a basketball into a hoop to show the parabolic curve it makes. (credit: modification of work by Dan Meyer)*

Answer available at the end of this section.

How To: Given a quadratic function in general form, find the vertex of the parabola.

1. Identify a, b, and c.
2. Find h, the x-coordinate of the vertex, by substituting a and b into $h = -\frac{b}{2a}$.
3. Find k, the y-coordinate of the vertex, by evaluating $k = f(h) = f\left(-\frac{b}{2a}\right)$.

Example 7.7.3: Finding the Vertex of a Quadratic Function

Find the vertex of the quadratic function $f(x) = 2x^2 – 6x + 7$. Rewrite the quadratic in standard form (vertex form).

Solution

The horizontal coordinate of the vertex will be at

$$h = -\frac{b}{2a} \tag{7.7.12}$$
$$= -\frac{-6}{2(2)} \tag{7.7.13}$$
$$= \frac{6}{4} \tag{7.7.14}$$
$$= \frac{3}{2} \tag{7.7.15}$$

The vertical coordinate of the vertex will be at

$$k = f(h) \tag{7.7.16}$$
$$= f\left(\frac{3}{2}\right) \tag{7.7.17}$$
$$= 2\left(\frac{3}{2}\right)^2 - 6\left(\frac{3}{2}\right) + 7 \tag{7.7.18}$$
$$= \frac{5}{2} \tag{7.7.19}$$

Rewriting into standard form, the stretch factor will be the same as the a in the original quadratic.

$$f(x) = ax^2 + bx + c \tag{7.7.20}$$
$$f(x) = 2x^2 - 6x + 7$$

Using the vertex to determine the shifts,

$$f(x) = 2\left(x - \frac{3}{2}\right)^2 + \frac{5}{2} \tag{7.7.21}$$

Analysis

One reason we may want to identify the vertex of the parabola is that this point will inform us what the maximum or minimum value of the function is, (k),and where it occurs, (h).

Exercise 7.7.3

Given the equation $g(x) = 13 + x^2 - 6x$, write the equation in general form and then in standard form.

Answer available at the end of this section.

Determining the Maximum and Minimum Values of Quadratic Functions

The output of the quadratic function at the vertex is the maximum or minimum value of the function, depending on the orientation of the **parabola**. We can see the maximum and minimum values in Figure 7.7.9.

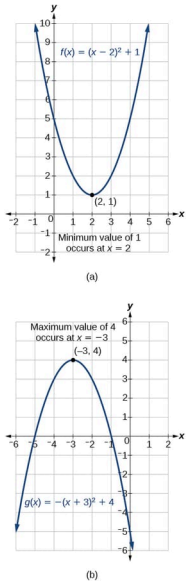


Figure 7.7.9: Minimum and maximum of two quadratic functions.

There are many real-world scenarios that involve finding the maximum or minimum value of a quadratic function, such as applications involving area and revenue.

Example 7.7.5: Finding the Maximum Value of a Quadratic Function

A backyard farmer wants to enclose a rectangular space for a new garden within her fenced backyard. She has purchased 80 feet of wire fencing to enclose three sides, and she will use a section of the backyard fence as the fourth side.

a. Find a formula for the area enclosed by the fence if the sides of fencing perpendicular to the existing fence have length L.
b. What dimensions should she make her garden to maximize the enclosed area?

Solution

Let's use a diagram such as Figure 7.7.10 to record the given information. It is also helpful to introduce a temporary variable, W, to represent the width of the garden and the length of the fence section parallel to the backyard fence.

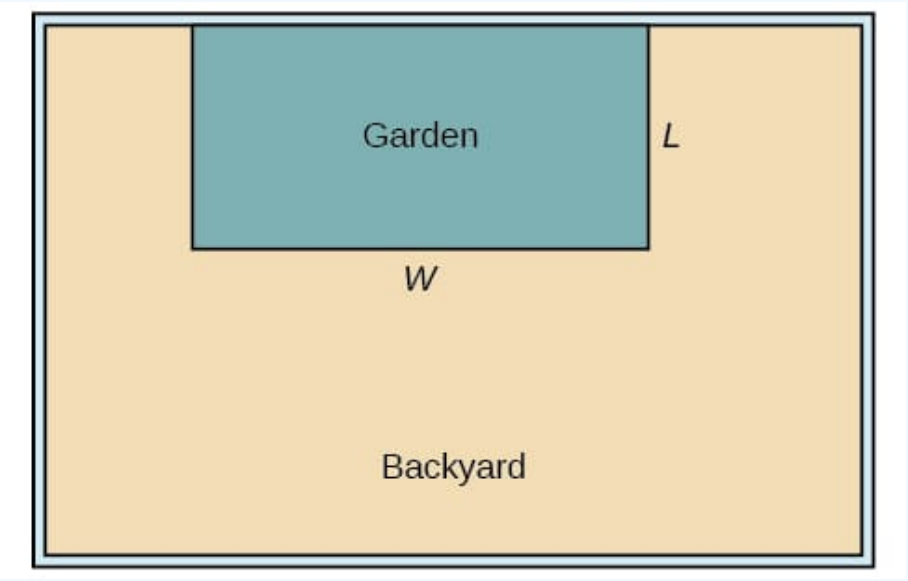


Figure 7.7.10: Diagram of the garden and the backyard.

a. We know we have only 80 feet of fence available, and $L + W + L = 80$, or more simply, $2L + W = 80$. This allows us to represent the width, W, in terms of L.

$$W = 80 - 2L \tag{7.7.22}$$

Now we are ready to write an equation for the area the fence encloses. We know the area of a rectangle is length multiplied by width, so

$$A = LW = L(80 - 2L) \tag{7.7.23}$$
$$A(L) = 80L - 2L^2 \tag{7.7.24}$$

This formula represents the area of the fence in terms of the variable length L. The function, written in general form, is

$$A(L) = -2L^2 + 80L \tag{7.7.25}$$

.

The quadratic has a negative leading coefficient, so the graph will open downward, and the vertex will be the maximum value for the area. In finding the vertex, we must be careful because the equation is not written in standard polynomial form with decreasing powers. This is why we rewrote the function in general form above. Since a is the coefficient of the squared term, $a = -2$, $b = 80$, and $c = 0$.

To find the vertex:

$$h = -\frac{80}{2(-2)} \qquad k = A(20) \tag{7.7.26}$$
$$= 20 \qquad \text{and} \qquad = 80(20) - 2(20)^2 \tag{7.7.27}$$
$$= 800 \tag{7.7.28}$$

The maximum value of the function is an area of 800 square feet, which occurs when $L = 20$ feet. When the shorter sides are 20 feet, there is 40 feet of fencing left for the longer side. To maximize the area, she should enclose the garden so the two shorter sides have length 20 feet and the longer side parallel to the existing fence has length 40 feet.

Analysis

This problem also could be solved by graphing the quadratic function. We can see where the maximum area occurs on a graph of the quadratic function in Figure 7.7.11.

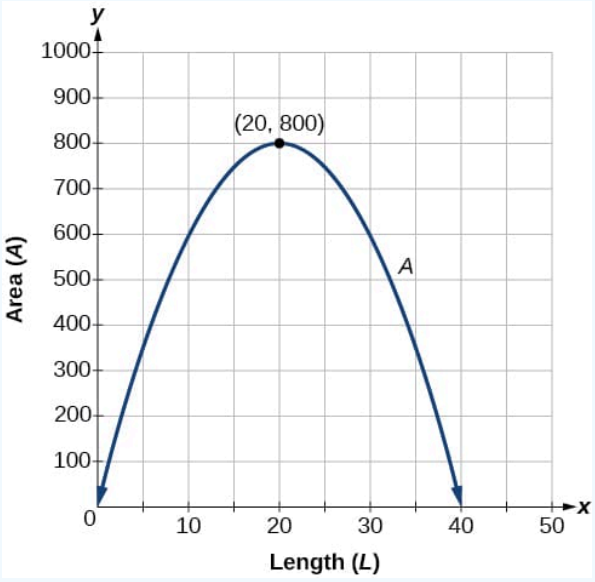


Figure 7.7.11: Graph of the parabolic function $A(L) = -2L^2 + 80L$

How To: Given an application involving revenue, use a quadratic equation to find the maximum.

1. Write a quadratic equation for revenue.
2. Find the vertex of the quadratic equation.
3. Determine the y-value of the vertex.

Example 7.7.6: Finding Maximum Revenue

The unit price of an item affects its supply and demand. That is, if the unit price goes up, the demand for the item will usually decrease. For example, a local newspaper currently has 84,000 subscribers at a quarterly charge of $30. Market research has suggested that if the owners raise the price to $32, they would lose 5,000 subscribers. Assuming that subscriptions are linearly related to the price, what price should the newspaper charge for a quarterly subscription to maximize their revenue?

Solution

Revenue is the amount of money a company brings in. In this case, the revenue can be found by multiplying the price per subscription times the number of subscribers, or quantity. We can introduce variables, p for price per subscription and Q for quantity, giving us the equation $\text{Revenue} = pQ$.

Because the number of subscribers changes with the price, we need to find a relationship between the variables. We know that currently $p = 30$ and $Q = 84,000$. We also know that if the price rises to $32, the newspaper would lose 5,000 subscribers, giving a second pair of values, $p = 32$ and $Q = 79,000$. From this we can find a linear equation relating the two quantities. The slope will be

$$m = \frac{79,000 - 84,000}{32 - 30} \tag{7.7.29}$$
$$= -\frac{5,000}{2} \tag{7.7.30}$$
$$= -2,500 \tag{7.7.31}$$

This tells us the paper will lose 2,500 subscribers for each dollar they raise the price. We can then solve for the y-intercept.

$$Q = -2500p + b \qquad \text{Substitute in the point } Q = 84,000 \text{ and } p = 30 \tag{7.7.32}$$
$$84,000 = -2500(30) + b \qquad \text{Solve for } b \tag{7.7.33}$$
$$b = 159,000 \tag{7.7.34}$$

This gives us the linear equation $Q = -2,500p + 159,000$ relating cost and subscribers. We now return to our revenue equation.

$$\text{Revenue} = pQ \tag{7.7.35}$$
$$\text{Revenue} = p(-2,500p + 159,000) \tag{7.7.36}$$
$$\text{Revenue} = -2,500p^2 + 159,000p \tag{7.7.37}$$

We now have a quadratic function for revenue as a function of the subscription charge. To find the price that will maximize revenue for the newspaper, we can find the vertex.

$$h = -\frac{159,000}{2(-2,500)} \tag{7.7.38}$$
$$= 31.8 \tag{7.7.39}$$

The model tells us that the maximum revenue will occur if the newspaper charges $31.80 for a subscription. To find what the maximum revenue is, we evaluate the revenue function.

$$\text{maximum revenue} = -2,500(31.8)^2 + 159,000(31.8) \tag{7.7.40}$$
$$= 2,528,100 \tag{7.7.41}$$

Analysis

This could also be solved by graphing the quadratic as in Figure 7.7.12. We can see the maximum revenue on a graph of the quadratic function.

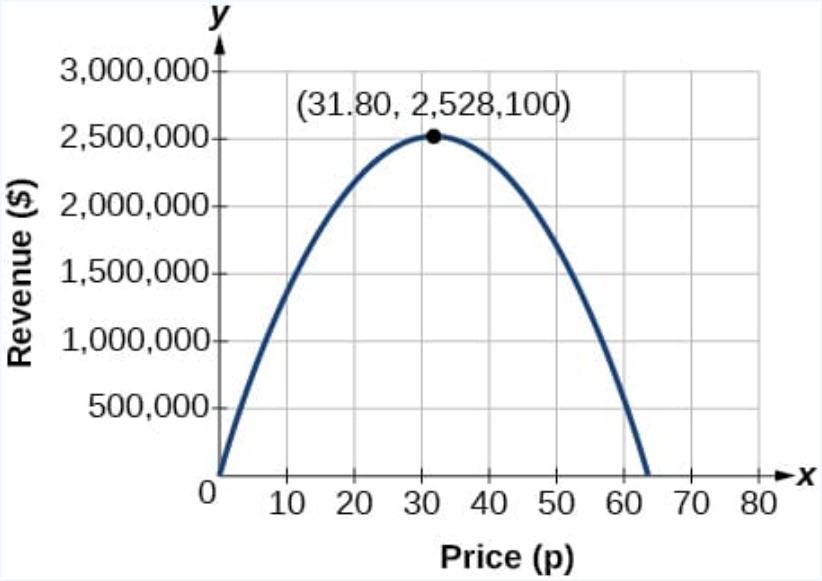


Figure 7.7.12: Graph of the parabolic function

Example 7.7.10: Applying the Vertex and x-Intercepts of a Parabola

A ball is thrown upward from the top of a 40 foot high building at a speed of 80 feet per second. The ball's height above ground can be modeled by the equation $H(t) = -16t^2 + 80t + 40$.

When does the ball reach the maximum height?
What is the maximum height of the ball?
When does the ball hit the ground?

The ball reaches the maximum height at the vertex of the parabola.

$$h = -\frac{80}{2(-16)} \tag{7.7.42}$$
$$= \frac{80}{32} \tag{7.7.43}$$
$$= \frac{5}{2} \tag{7.7.44}$$
$$= 2.5 \tag{7.7.45}$$

The ball reaches a maximum height after 2.5 seconds.

To find the maximum height, find the y-coordinate of the vertex of the parabola.

$$k = H(-\frac{b}{2a}) \tag{7.7.46}$$
$$= H(2.5) \tag{7.7.47}$$
$$= -16(2.5)^2 + 80(2.5) + 40 \tag{7.7.48}$$
$$= 140 \tag{7.7.49}$$

The ball reaches a maximum height of 140 feet.

To find when the ball hits the ground, we need to determine when the height is zero, $H(t) = 0$.

We use the quadratic formula.

$$t = \frac{-80 \pm \sqrt{80^2 - 4(-16)(40)}}{2(-16)} \tag{7.7.50}$$

$$= \frac{-80 \pm \sqrt{8960}}{-32} \tag{7.7.51}$$

Because the square root does not simplify nicely, we can use a calculator to approximate the values of the solutions.

$$t = \frac{-80 - \sqrt{8960}}{-32} \approx 5.458 \text{ or } t = \frac{-80 + \sqrt{8960}}{-32} \approx -0.458 \tag{7.7.52}$$

The second answer is outside the reasonable domain of our model, so we conclude the ball will hit the ground after about 5.458 seconds. See Figure 7.7.16.

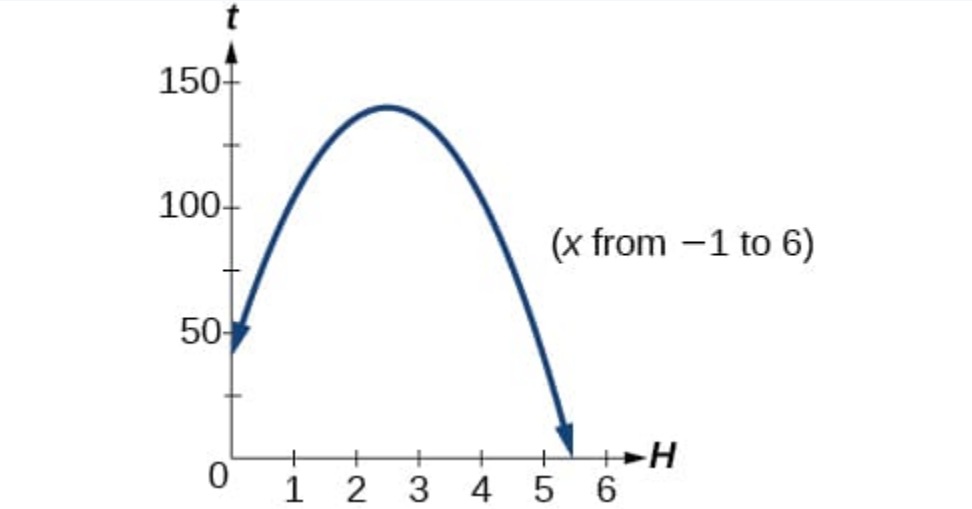


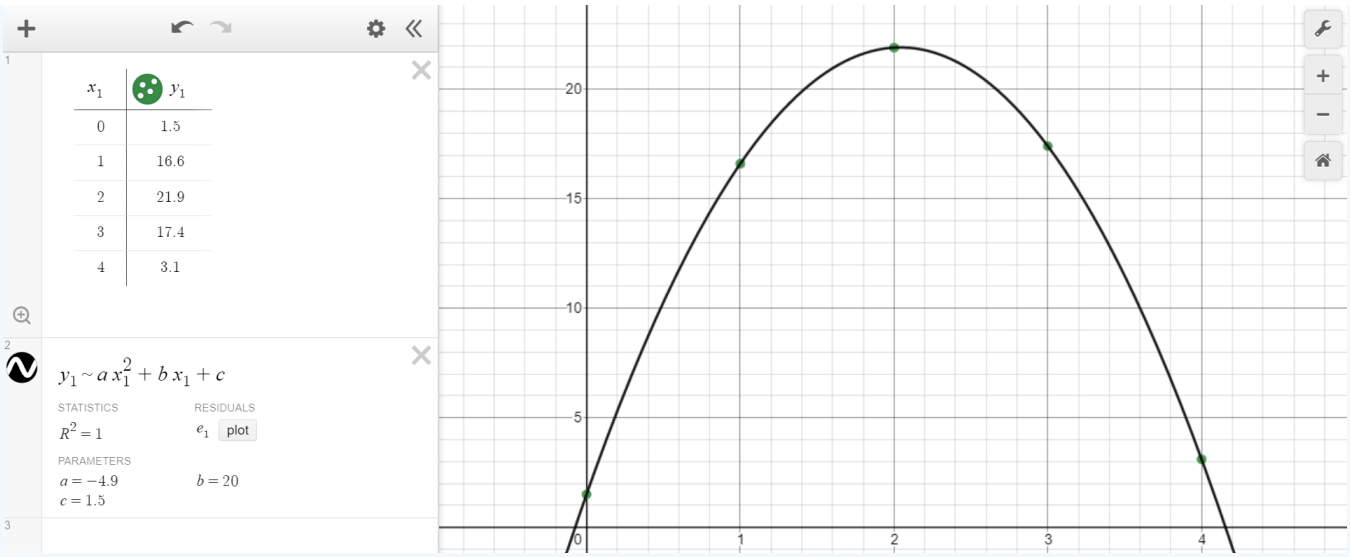
Figure 7.7.13

Example 7.7.11: Using Technology to Find the Best Fit Quadratic Model

A ball is thrown into the air, and the following data is collected where x represents the time in seconds after the ball is thrown up and y represents the height in meters of the ball. We can use desmos to create a quadratic model that fits the given data.

x	0	1	2	3	4
y	1.5	16.6	21.9	17.4	3.1

On desmos, type the data into a table with the x-values in the first column and the y-values in the second column. Then, to tell desmos to compute a quadratic model, type in $y_1 \sim a\, x_1^2 + b\, x_1 + c$. You will get a result that looks like this:



You can go to this problem in desmos by clicking https://www.desmos.com/calculator/u8ytorpnhk. Looking at the results, the quadratic model that fits the data is

$$y = -4.9x^2 + 20x + 1.5 \tag{7.7.53}$$

Answers

Exercise 7.7.2

The path passes through the origin and has vertex at $(-4, 7)$, so $h(x) = -\frac{7}{16}(x+4)^2 + 7$. To make the shot, $h(-7.5)$ would need to be about 4 but $h(-7.5) \approx 1.64$; he doesn't make it.

Exercise 7.7.3

$g(x) = x^2 - 6x + 13$ in general form; $g(x) = (x-3)^2 + 4$ in standard form.

This page titled 7.7: Modeling with Quadratic Functions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 3.2: Quadratic Functions by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

7.8: Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables *x* and *y*. The most common and easiest way is a *scatter plot*. The following example illustrates a scatter plot.

Example 7.8.1

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let $x =$ the year and let $y =$ the number of m-commerce users, in millions.

Table 7.8.1: Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

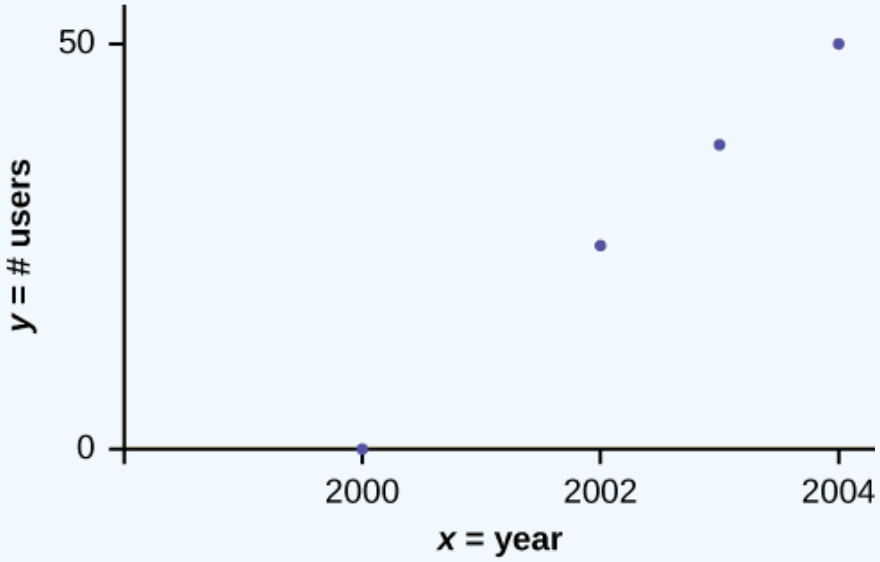


Figure 7.8.1: Scatter plot showing the number of m-commerce users (in millions) by year.

To create a scatter plot

a. Enter your X data into list L1 and your Y data into list L2.
b. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
c. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
d. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
e. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
f. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.

g. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Exercise 7.8.1

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Answer available at the end of this section.

A scatter plot shows the *direction of a relationship* between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the *strength of the relationship* by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatter plot, you want to notice the *overall pattern* and any *deviations* from the pattern. The following scatterplot examples illustrate these concepts.

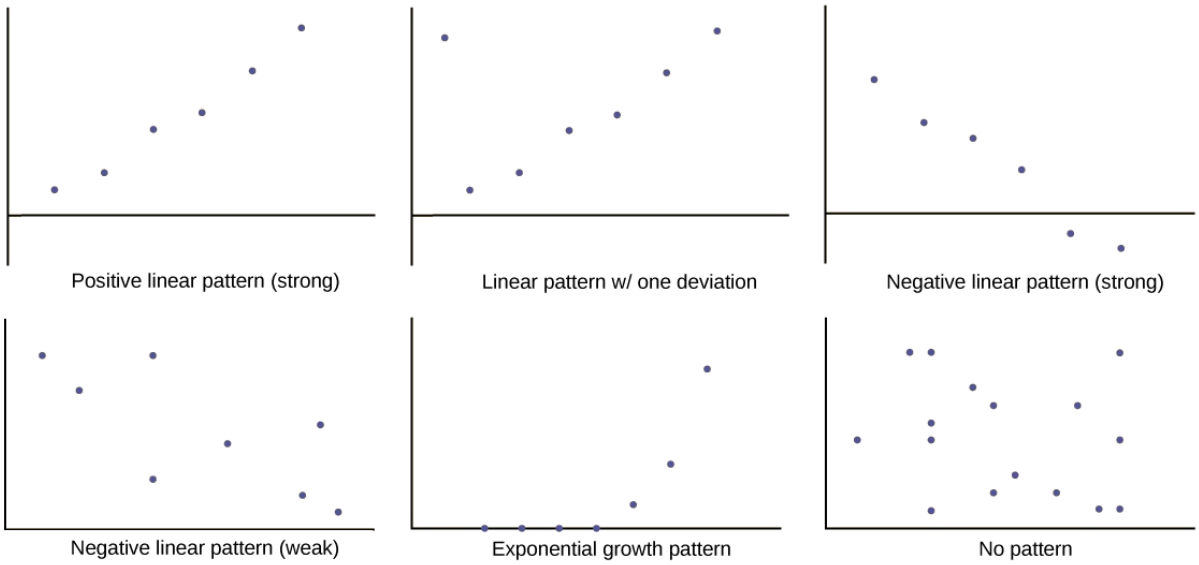


Figure 7.8.3:

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

Summary

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

Contributors and Attributions

- Template:ContribOpenStax

Answers

Exercise 7.8.1

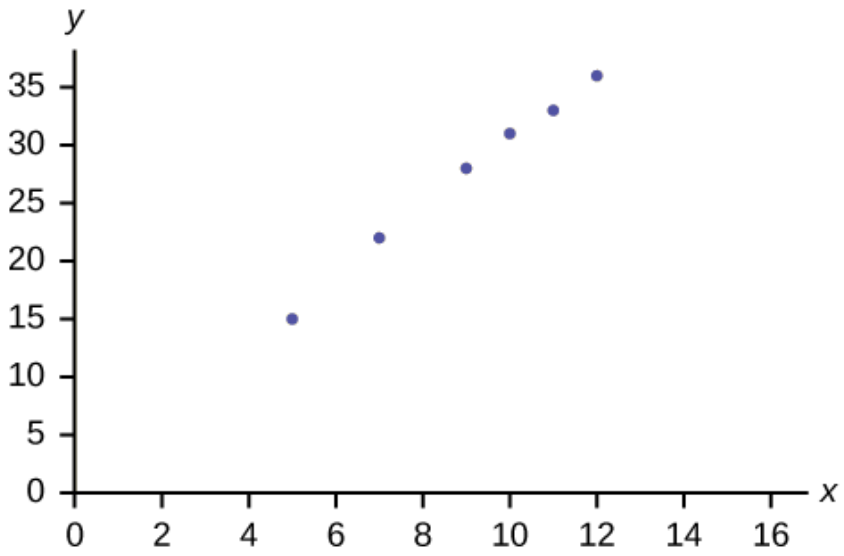


Figure 7.8.2

Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

This page titled 7.8: Scatter Plots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

8: Additional Topics

Close message

Learn more

Welcome to your new guide!

Click on "New" to build out your guide.

- Topic hierarchy

This page titled 8: Additional Topics is shared under a license and was authored, remixed, and/or curated by LibreTexts.

INDEX

A

Absolute Value Function
7.1: Introduction to Modeling
Aconstant rate of change
7.2: Modeling with Linear Functions
ad hominem
3.6: Common Valid and Invalid Arguments
algebraic expression
1.1: Numbers and Operations
1.3: Algebra Essentials
appeal to authority
3.6: Common Valid and Invalid Arguments
appeal to consequence
3.6: Common Valid and Invalid Arguments
appeal to ignorance
3.6: Common Valid and Invalid Arguments
argument
3.4: Inductive and Deductive Reasoning
associative property of addition
1.1: Numbers and Operations
1.3: Algebra Essentials
associative property of multiplication
1.1: Numbers and Operations
1.3: Algebra Essentials
axis of symmetry
7.7: Modeling with Quadratic Functions

B

base
1.1: Numbers and Operations
1.3: Algebra Essentials
binomial probability distribution
6.5: Estimating Population Proportion

C

Cartesian coordinate plane
1.4: Graphing
circular reasoning
3.6: Common Valid and Invalid Arguments
cluster sampling
4.2: Collecting Data
combinations
5.4: Counting Methods
commutative property of addition
1.1: Numbers and Operations
1.3: Algebra Essentials
complement
3.2: Operations with Sets
composite function
7.1: Introduction to Modeling
conditional probability
5.5: Conditional Probabilities
constant
1.1: Numbers and Operations
1.3: Algebra Essentials
constant rate of change
7.2: Modeling with Linear Functions
continuous data
4.2: Collecting Data
correlation coefficient
7.3: Fitting Linear Models to Data
correlation implies causation
3.6: Common Valid and Invalid Arguments
cumulative probability distributions
6.2: Continuous Random Variables
cumulative relative frequency
4.3: Visual Summaries of Data

D

decreasing on the interval I
7.1: Introduction to Modeling
deductive
3.4: Inductive and Deductive Reasoning
dependent variable
7.1: Introduction to Modeling
direction of a relationship between the variables
7.8: Scatter Plots
discrete data
4.2: Collecting Data
Distributive Property
1.1: Numbers and Operations
1.3: Algebra Essentials
domain
7.1: Introduction to Modeling

E

elements
3.1: The Basics
empty set
3.1: The Basics
equation
1.1: Numbers and Operations
1.3: Algebra Essentials
even function
7.1: Introduction to Modeling
expected value
5.2: Expected Value
exponent
1.1: Numbers and Operations
1.3: Algebra Essentials
exponential notation
1.1: Numbers and Operations
1.3: Algebra Essentials
exponential regression
7.5: Fitting Exponential Models to Data
extrapolation
7.3: Fitting Linear Models to Data

F

factorial
5.4: Counting Methods
fair game
5.2: Expected Value
false dilemma
3.6: Common Valid and Invalid Arguments
formula
1.1: Numbers and Operations
1.3: Algebra Essentials
frequency
4.3: Visual Summaries of Data
Frequency Polygons
4.4: Other Summaries
frequency table
4.3: Visual Summaries of Data
function
7.1: Introduction to Modeling

G

Gambler's Fallacy
5.2: Expected Value
general form of a quadratic function
7.7: Modeling with Quadratic Functions
graph of a function
7.1: Introduction to Modeling

H

histograms
4.4: Other Summaries
house edge
5.2: Expected Value

I

identity property of addition
1.1: Numbers and Operations
1.3: Algebra Essentials
identity property of multiplication
1.1: Numbers and Operations
1.3: Algebra Essentials
increasing on the interval I
7.1: Introduction to Modeling
independent variable
7.1: Introduction to Modeling
inductive
3.4: Inductive and Deductive Reasoning
inferential statistics
6.4: Estimating Population Mean
integers
1.1: Numbers and Operations
1.3: Algebra Essentials
interpolation
7.3: Fitting Linear Models to Data
intersection
3.2: Operations with Sets
inverse property of addition
1.1: Numbers and Operations
1.3: Algebra Essentials
inverse property of multiplication
1.1: Numbers and Operations
1.3: Algebra Essentials
irrational numbers
1.1: Numbers and Operations
1.3: Algebra Essentials

L

law of large numbers
5.2: Expected Value

least squares regression
7.3: Fitting Linear Models to Data

level of measurement
4.3: Visual Summaries of Data

linear function
7.2: Modeling with Linear Functions

linear model
7.3: Fitting Linear Models to Data

linear model,
7.3: Fitting Linear Models to Data

linear relationship
7.3: Fitting Linear Models to Data

logarithmic regression
7.5: Fitting Exponential Models to Data

logical fallacies
3.6: Common Valid and Invalid Arguments

logistic regression
7.5: Fitting Exponential Models to Data

M

median
4.5: Measures of the Center and Variation of the Data
4.6: Measures of the Location of the Data and Outliers

mode
4.5: Measures of the Center and Variation of the Data

model breakdown
7.3: Fitting Linear Models to Data

modeling
7.2: Modeling with Linear Functions

multiplication tree
5.4: Counting Methods

N

natural numbers
1.1: Numbers and Operations
1.3: Algebra Essentials

notation
3.1: The Basics

O

odd function
7.1: Introduction to Modeling

order of operations
1.1: Numbers and Operations
1.3: Algebra Essentials

ordered pairs
1.4: Graphing

outliers
4.6: Measures of the Location of the Data and Outliers

P

parameter
4.1: Introduction to Statistics

Pareto chart
4.2: Collecting Data

permutation
5.4: Counting Methods

population
4.1: Introduction to Statistics

population mean
4.5: Measures of the Center and Variation of the Data

post hoc
3.6: Common Valid and Invalid Arguments

principal
2.2: Simple Interest

probability
4.1: Introduction to Statistics

probability distribution function
6.1: Discrete Random Variables

proper subset
3.1: The Basics

Q

qualitative data
4.2: Collecting Data

quantitative data
4.2: Collecting Data

quartiles
4.6: Measures of the Location of the Data and Outliers

R

range
7.1: Introduction to Modeling

rate of change is constant
7.2: Modeling with Linear Functions

rational numbers
1.1: Numbers and Operations
1.3: Algebra Essentials

real number line
1.1: Numbers and Operations
1.3: Algebra Essentials

real numbers
1.1: Numbers and Operations
1.3: Algebra Essentials

regression
7.5: Fitting Exponential Models to Data

rounding
4.3: Visual Summaries of Data

S

sample mean
4.5: Measures of the Center and Variation of the Data

sampling bias
4.2: Collecting Data

Sampling Error
4.2: Collecting Data

sampling with replacement
4.2: Collecting Data

sampling without replacement
4.2: Collecting Data

scatter plot
7.8: Scatter Plots

sets
3.1: The Basics

simple interest
2.2: Simple Interest

standard form of a quadratic function
7.7: Modeling with Quadratic Functions

standard normal distribution
6.3: Normal Distributions

statistic
4.1: Introduction to Statistics

straw man
3.6: Common Valid and Invalid Arguments

strength of a relationship between the variables
7.8: Scatter Plots

subset
3.1: The Basics

syllogism
3.4: Inductive and Deductive Reasoning

symmetry about the origin
7.1: Introduction to Modeling

T

table of values
7.1: Introduction to Modeling

Time Series Graphs
4.4: Other Summaries

U

unions
3.2: Operations with Sets

universal
3.2: Operations with Sets

V

variable
1.1: Numbers and Operations
1.3: Algebra Essentials
4.1: Introduction to Statistics

Venn diagram
3.3: Venn Diagrams

vertex
7.7: Modeling with Quadratic Functions

vertex form of a quadratic function
7.7: Modeling with Quadratic Functions

vertical line test
7.1: Introduction to Modeling

W

whole numbers
1.1: Numbers and Operations
1.3: Algebra Essentials

Z

zeros
7.7: Modeling with Quadratic Functions

zeros of a function
7.1: Introduction to Modeling

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

GLOSSARY

S

Sample Word 1 | Sample Definition 1

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

DETAILED LICENSING

Overview

Title: Ideas of Mathematics

Webpages: 58

Applicable Restrictions: Noncommercial

All licenses found:

- Undeclared: 36.2% (21 pages)
- CC BY 4.0: 34.5% (20 pages)
- CC BY-SA 4.0: 15.5% (9 pages)
- CC BY-NC-SA 4.0: 13.8% (8 pages)

By Page

- Ideas of Mathematics - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Number Sense - *Undeclared*
    - 1.1: Numbers and Operations - *CC BY 4.0*
    - 1.2: Percent - *CC BY-NC-SA 4.0*
    - 1.3: Algebra Essentials - *CC BY 4.0*
    - 1.4: Graphing - *CC BY-NC-SA 4.0*
  - 2: Finance - *Undeclared*
    - 2.1: Budgeting - *Undeclared*
    - 2.2: Simple Interest - *CC BY-SA 4.0*
    - 2.3: Compound Interest - *CC BY-SA 4.0*
    - 2.4: Installments - *CC BY-SA 4.0*
  - 3: Set Theory and Logic - *Undeclared*
    - 3.1: The Basics - *CC BY-NC-SA 4.0*
    - 3.2: Operations with Sets - *CC BY-NC-SA 4.0*
    - 3.3: Venn Diagrams - *CC BY-NC-SA 4.0*
    - 3.4: Inductive and Deductive Reasoning - *CC BY-NC-SA 4.0*
    - 3.5: Using Venn Diagrams to Analyze Arguments - *Undeclared*
    - 3.6: Common Valid and Invalid Arguments - *CC BY-NC-SA 4.0*
    - 3.7: Truth Tables - *CC BY 4.0*
    - 3.8: Analyzing Arguments with Truth Tables - *CC BY 4.0*
  - 4: Descriptive Statistics - *Undeclared*
    - 4.1: Introduction to Statistics - *CC BY 4.0*
    - 4.2: Collecting Data - *CC BY 4.0*
    - 4.3: Visual Summaries of Data - *CC BY 4.0*
    - 4.4: Other Summaries - *CC BY 4.0*
    - 4.5: Measures of the Center and Variation of the Data - *CC BY 4.0*
    - 4.6: Measures of the Location of the Data and Outliers - *CC BY 4.0*
  - 5: Probability - *Undeclared*
    - 5.1: Introduction on Probability - *CC BY-SA 4.0*
    - 5.2: Expected Value - *CC BY-SA 4.0*
    - 5.3: Probability Rules- “And” and “Or” - *CC BY-SA 4.0*
    - 5.4: Counting Methods - *CC BY-SA 4.0*
    - 5.5: Conditional Probabilities - *CC BY-SA 4.0*
  - 6: Inferential Statistics - *Undeclared*
    - 6.1: Discrete Random Variables - *CC BY 4.0*
    - 6.2: Continuous Random Variables - *CC BY 4.0*
    - 6.3: Normal Distributions - *CC BY 4.0*
    - 6.4: Estimating Population Mean - *CC BY 4.0*
    - 6.5: Estimating Population Proportion - *CC BY 4.0*
  - 7: Modeling - *Undeclared*
    - 7.1: Introduction to Modeling - *CC BY-NC-SA 4.0*
    - 7.2: Modeling with Linear Functions - *CC BY 4.0*
    - 7.3: Fitting Linear Models to Data - *CC BY 4.0*
    - 7.4: Modeling with Exponential Functions - *CC BY-SA 4.0*
    - 7.5: Fitting Exponential Models to Data - *CC BY 4.0*
    - 7.6: Putting It All Together - *Undeclared*
    - 7.7: Modeling with Quadratic Functions - *CC BY 4.0*
    - 7.8: Scatter Plots - *CC BY 4.0*
  - 8: Additional Topics - *Undeclared*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*

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